(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **2002170891 A**

(43) Date of publication of application: **14.06.02**

(51) Int. Cl **H01L 21/8247**
H01L 29/788
H01L 29/792
H01L 27/115

(21) Application number: **2000354722**

(22) Date of filing: **21.11.00**

(71) Applicant: **HALO LSI DESIGN & DEVICE TECHNOL INC NEW HEIRO:KK**

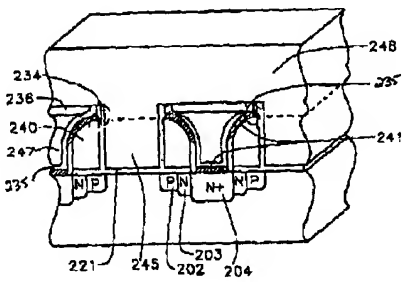
(72) Inventor: **OGURA SEIKI**
OGURA TOMOKO
HAYASHI YUTAKA

(54) **MANUFACTURE OF DUAL BIT MULTI-LEVEL BALLISTIC MONOS MEMORY, PROGRAMMING, AND OPERATION PROCESS**

(57) Abstract:

PROBLEM TO BE SOLVED: To provide a flash memory of fast low-voltage ballistic program, ultra-short channel, ultra-high integration level, and dual bit multi-level.

SOLUTION: A cell structure is realized by (i) providing a side wall control gate on the laminated film of oxide film, nitride film, oxide film (ONO) on both sides of a ward gate, and (ii) forming a control gate and a bit impurity film by self-alignment so that the control gate and the bit impurity film are shared between adjoining memory cells due to high integration. The process comprises, as main components, 1) a process for manufacturing a removable side wall for manufacturing the ultra-short channel and the side wall control gate with or without a step structure, and 2) the formation of the control gate, by self-alignment, on a storage nitride film and the impurity film.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号
特開2002－170891
（P2002－170891A）
(43)公開日　平成14年６月14日(2002.6.14)

(51)Int.Cl.[7]	識別記号	FI		テーマコード（参考）
H01L 21/8247		H01L 29/78	371	5F001
29/788			434	5F083
29/792				5F101
27/115				

審査請求　未請求　請求項の数86　ＯＬ　外国語出願　（全 74 頁）

(21)出願番号　特願2000－354722(P2000－354722)

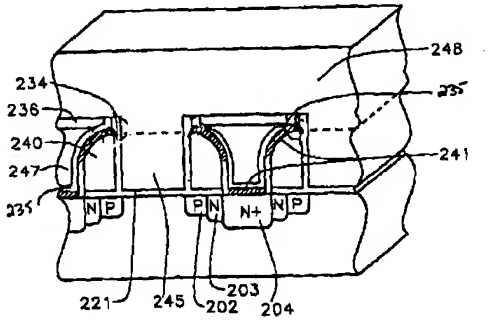
(22)出願日　平成12年11月21日(2000.11.21)

(71)出願人　599154261
ヘイロ　エルエスアイ　デザインアンドディヴァイス　テクノロジー　インコーポレイテッド
アメリカ合衆国　12590　ニューヨーク州、ワッピンガーズ　フォールズ、メイヤーズ　コーナーズ　ロード　169
(71)出願人　500361799
株式会社ニューヘイロ
東京都杉並区高井戸東３丁目２番24号
(74)代理人　100084870
弁理士　田中　香樹　（外１名）

最終頁に続く

(54)【発明の名称】　デュアルビット多準位バリスティックMONOSメモリの製造、プログラミング、および動作のプロセス

(57)【要約】　（修正有）
【課題】高速低電圧バリスティックプログラム、超短チャネル、超高集積度、デュアルビット多準位のフラッシュメモリを提供する。
【解決手段】セル構造は、（ｉ）ワードゲートの両サイド上の酸化膜－窒化膜－酸化膜（ＯＮＯ）の積層膜上にサイドウォール制御ゲートを配設すること、および（ｉｉ）自己整合によって制御ゲートおよびビット不純膜を形成し、高集積のために隣接するメモリセル間の制御ゲートおよびビット不純膜を共有することによって実現される。本プロセスで用いられる主要素は、１）ステップ構造を有するか、または無しで、超短チャネルおよびサイドウォール制御ゲートを製造するための、除去可能なサイドウォールの製造プロセス、および２）蓄積窒化膜および不純物膜上の制御ゲートの自己整合による形成である。



【特許請求の範囲】

【請求項１】　半導体基板の表面上にゲートシリコン酸化膜を形成すること、
前記ゲートシリコン酸化膜を覆うように第１のポリシリコン膜を堆積すること、
前記第１ポリシリコン膜を覆うように第１の窒化膜を堆積すること、
ワードゲートが、その間に間隙が残るように形成されるように、前記第１のポリシリコン膜および前記第１の窒化膜をパターニングすること、
前記ワードゲートのサイドウォール上に第１の絶縁膜を形成すること、
前記ワードゲートおよび前記ゲートシリコン酸化膜を覆うようにスペーサ膜を堆積すること、
除去可能なスペーサが前記ワードゲートのサイドウォール上に残るように、前記スペーサ膜を異方性エッチングにより除去すること、
浅いドープ領域を形成するために、前記除去可能なスペーサをイオン打ち込みマスクとして機能させて、前記半導体基板内にイオンを打ち込むこと、
その後に前記除去可能なスペーサを取り除くこと、
前記間隙内の前記半導体基板上に窒化物含有膜を堆積すること、
前記ワードゲートおよび前記窒化物含有膜上に第２のポリシリコン膜を堆積すること、
前記ワードゲートのサイドウォール上に、制御サイドウォールスペーサゲートとなるポリシリコンスペーサが残り、電荷が蓄積される窒化物領域を形成する窒化物含有膜が前記制御サイドウォールスペーサゲートのそれぞれの下に形成されるように、前記第２のポリシリコン膜および前記窒化物含有膜を異方性エッチングにより除去すること、
前記制御サイドウォールスペーサゲート上に第２の絶縁膜を形成すること、ビット拡散領域を形成するために、前記制御サイドウォールスペーサゲートを打ち込みマスクとして機能させて、前記半導体基板内にイオンを打ち込むこと、
前記２つのサイドワードゲート間の間隙を充填するような間隙充填材で前記基板の表面を被覆すること、
前記間隙充填材料を平坦にすること、
その後に前記ワードゲート上の前記第１の窒化膜を除去すること、およびＭＯＮＯＳメモリ素子の前記製造を完了させるために、前記ワードゲート下で接続されるワード線を形成するような第３のポリシリコン膜を、前記基板上に堆積することを含むＭＯＮＯＳメモリデバイスの製造方法。

【請求項２】　前記ゲートシリコン酸化膜が約５～１０ｎｍの厚さを有する請求項１の製造方法。

【請求項３】　前記第１のポリシリコン膜がＣＶＤにより約１５０～２５０ｎｍの厚さにまで堆積される請求項１の製造方法。

【請求項４】　前記第１の窒化膜がＣＶＤにより約５０～１００ｎｍの厚さにまで堆積される請求項１の製造方法。

【請求項５】　前記第１の絶縁膜が、前記ワードゲートのサイドウォールの表面を熱酸化して約５～１０ｎｍの厚さに形成されたシリコン酸化膜である請求項１の製造方法。

【請求項６】　前記第１の絶縁膜が、ＣＶＤによって前記ワードゲートのサイドウォール上に約５～１０ｎｍの厚さに堆積されたシリコン酸化膜である請求項１の製造方法。

【請求項７】　前記第１の絶縁膜が、前記ワードゲートのサイドウォール上に約５～１０ｎｍの厚さに堆積されたシリコン窒化膜である請求項１の製造方法。

【請求項８】　前記第１の絶縁膜が、前記ワードゲートのサイドウォール上に、合わせて約５～１０ｎｍの厚さに堆積されたシリコン酸化膜およびシリコン窒化膜である請求項１の製造方法。

【請求項９】　前記スペーサ膜が、ポリシリコン、プラズマ窒化膜、プラズマ酸化窒化膜、およびホウ素リンガラス（ＢＰＳＧ）を含むグループ内のいずれかによって構成され、約３０～５０ｎｍの厚さである請求項１の製造方法。

【請求項１０】　前記除去可能なスペーサを除去する段階が乾式の化学的異法性エッチングを含む請求項１の製造方法。

【請求項１１】　前記窒化物含有膜を堆積する段階が、前記半導体基板上に、約３．６～５．０ｎｍの厚さにまで第１のシリコン酸化膜を成長させること、
前記第１のシリコン酸化膜上に、約２～５ｎｍの厚さを有するシリコン窒化膜を堆積すること、および前記シリコン窒化膜上に、約４～８ｎｍの厚さを有する第２のシリコン酸化膜を堆積することを含む請求項１の製造方法。

【請求項１２】　前記シリコン窒化膜を堆積する段階の前に、前記第１のシリコン酸化膜を窒化することをさらに含む請求項１の製造方法。

【請求項１３】　前記第２のポリシリコン膜が約３０～５０ｎｍの厚さを有する請求項１の製造方法。

【請求項１４】　前記第２のポリシリコン膜が約３０～５０ｎｍの厚さを有し、約６０～１００ｎｍの厚さを有するタングステンシリサイド層を堆積することをさらに含み、前記第２のポリシリコン膜およびタングステンシリサイド層が共に前記制御サイドウォールスペーサゲートを形成する請求項１の製造方法。

【請求項１５】　前期第２の絶縁膜が、ＣＶＤによって約１０ｎｍの厚さにまで堆積されたシリコン酸化膜を含む請求項１の製造方法。

【請求項１６】　前記第２の絶縁膜が、ＣＶＤによって

約１０ｎｍの厚さにまで堆積されたシリコン窒化膜を含む請求項１の製造方法。

【請求項１７】　前記制御サイドウォールスペーサゲートの下層部分にサイドウォール酸化膜スペーサを形成するために、前記第２の絶縁膜に異方性エッチングを施すこと、およびその後に、前記制御サイドスペーサゲートの上層部分および前記ビット拡散領域をシリサイド化することをさらに含む請求項１の製造方法。

【請求項１８】　前記間隙充填材が、シリコン酸化膜およびホウ素リンガラス（ＢＰＳＧ）を含むグループのいずれかにより構成される請求項１の製造方法。

【請求項１９】　前記間隙充填材が伝導性材料を含み、さらに前記伝導性材料を前記第１の窒化膜の表面下にへこませること、

前記へこまされた伝導性材料上にシリコン酸化膜を堆積すること、および前記シリコン酸化膜を平坦にすることを含み、

前記伝導性材料および下層の前記制御サイドウォールスペーサゲートが共に制御ゲートを形成する請求項１の製造方法。

【請求項２０】　前記第３のポリシリコン膜が約１５０～２００ｎｍの厚さを有する請求項１の製造方法。

【請求項２１】　前記ワード線をシリサイド化することをさらに含む請求項１の製造方法。

【請求項２２】　半導体基板の表面上にゲートシリコン酸化膜を形成すること、

前記ゲートシリコン酸化膜を覆うように第１のポリシリコン膜を堆積すること、

前記第１ポリシリコン膜を覆うように第１の窒化膜を堆積すること、

ワードゲートが、その間に間隙が残るように形成されるように、前記第１のポリシリコン膜および前記第１の窒化膜をパターニングすること、

前記ワードゲートのサイドウォール上に第１の絶縁膜を形成すること、

前記ワードゲートおよび前記ゲートシリコン酸化膜を覆うようにスペーサ膜を堆積すること、

除去可能なスペーサが前記ワードゲートのサイドウォール上に残るように、前記スペーサ膜を異方性エッチングにより除去すること、

前記半導体基板の１部を露出するために、前記ワードゲートおよび前記除去可能なスペーサによって覆われない部分の前記ゲートシリコン酸化膜をエッチングすること、

前記半導体基板の露出部分をエッチングすることにより、前記基板に段差を形成すること、

浅いドープ領域を形成するために、前記除去可能なスペーサをイオン打ち込みマスクとして機能させて、前記半導体基板内にイオンを打ち込むこと、

その後に前記除去可能なスペーサを取り除くこと、

前記除去可能なポリシリコンスペーサ下のゲートシリコン酸化膜を除去すること、

前記半導体基板上に酸化膜－窒化膜－酸化膜の積層膜を形成すること、

前記ワードゲートおよび前記第２のゲートシリコン酸化膜上に第２のポリシリコン膜を堆積すること、

前記ワードゲートのサイドウォール上に、サイドウォール制御ゲートとなるポリシリコンスペーサが残り、電荷が蓄積される窒化物領域を形成する酸化膜－窒化膜－酸化膜の積層膜の窒化部分が前記サイドウォール制御ゲートのそれぞれの下に形成されるように、前記第２のポリシリコン膜および前記酸化膜－窒化膜－酸化膜の積層膜を異方性エッチングにより除去すること、

前記制御サイドウォールスペーサゲート上に第２の絶縁膜を形成すること、

ビット拡散領域を形成するために、前記制御サイドウォールゲートを打ち込みマスクとして機能させて、前記半導体基板内にイオンを打ち込むこと、

前記２つのサイドワードゲート間の間隙を充填するような間隙充填材で前記基板の表面を被覆すること、

前記間隙充填材料を平坦にすること、

その後に前記ワードゲート上の前記第１の窒化膜を除去すること、およびＭＯＮＯＳメモリ素子の前記製造を完了させるために、前記ワードゲート下で接続されるワード線を形成するような第３のポリシリコン膜を、前記基板上に堆積することを含むステップスピリット構造ＭＯＮＯＳメモリデバイスの製造方法。

【請求項２３】　前記第１のポリシリコン膜がＣＶＤにより約１５０～２５０ｎｍの厚さにまで堆積される請求項２２の製造方法。

【請求項２４】　前記第１の窒化膜がＣＶＤにより約５０～１００ｎｍの厚さにまで堆積される請求項２２の製造方法。

【請求項２５】　前記第１の絶縁膜が、前記ワードゲートのサイドウォールの表面を熱酸化して約５～１０ｎｍの厚さに形成されたシリコン酸化膜である請求項２２の製造方法。

【請求項２６】　前記ワードゲートのサイドウォール上の第１の絶縁膜が約５～１０ｎｍの厚さである請求項２２の製造方法。

【請求項２７】　前記スペーサ膜が、ポリシリコン、プラズマ窒化膜、プラズマ酸化窒化膜、およびホウ素リンガラス（ＢＰＳＧ）を含むグループ内のいずれかによって構成され、約３０～５０ｎｍの厚さである請求項２２の製造方法。

【請求項２８】　前記除去可能なスペーサを除去する段階が乾式の化学的異法性エッチングを含む請求項２２の製造方法。

【請求項２９】　前記半導体基板に形成される段差が約２０～５０ｎｍの深さを有する請求項２２の製造方法。

【請求項３０】　前記除去可能なスペーサ下の前記ゲートシリコン酸化膜を除去する工程の後に、前記段差の角を丸めることをさらに含む請求項２２の製造方法。
【請求項３１】　前記段差の角の丸め工程が、約１０００～１１００℃で、約６０秒間の高速熱焼きなましを含む請求項３０の製造方法。
【請求項３２】　前記段差の角の丸め工程が、約９００℃、約２００～３００mトール圧の水素内での焼きなましを含む請求項３０の製造方法。
【請求項３３】　前記酸化膜－窒化膜－酸化膜の積層膜が、
約3．6～5．0nmの厚さを有する第１のシリコン酸化膜、
約2～5nmの厚さを有する第2のシリコン窒化膜、および約4～8nmの厚さを有する第3のシリコン酸化膜を含む請求項２２の製造方法。
【請求項３４】　前記第2のポリシリコン膜が約３０～５０nmの厚さを有する請求項２２の製造方法。
【請求項３５】　前記第2のポリシリコン膜が約３０～５０nmの厚さを有し、約６０～１００nmの厚さを有するタングステンシリサイド層を堆積することをさらに含み、前記第3のポリシリコン膜およびタングステンシリサイド層が共に前記制御サイドウォールゲートを形成する請求項２２の製造方法。
【請求項３６】　前記第2の絶縁膜が、ＣＶＤによって約１０nmの厚さにまで堆積されたシリコン酸化膜を含む請求項２２の製造方法。
【請求項３７】　前記第2の絶縁膜が、ＣＶＤによって約１０nmの厚さにまで堆積されたシリコン窒化膜を含む請求項２２の製造方法。
【請求項３８】　前記制御サイドウォールスペーサゲートの下層部分にサイドウォール酸化膜スペーサを形成するために、前記第2の絶縁膜に異方性エッチングを施すこと、およびその後に、前記制御サイドスペーサゲートの上層部分および前記ビット拡散領域をシリサイド化することをさらに含む請求項２２の製造方法。
【請求項３９】　前記間隙充填材が、シリコン酸化膜およびホウ素リンガラス（ＢＰＳＧ）を含むグループのいずれかにより構成される請求項２２の製造方法。
【請求項４０】　前記間隙充填材が伝導性材料を含み、さらに前記伝導性材料を前記第１の窒化膜の表面下にへこませること、
前記へこまされた伝導性材料上にシリコン酸化膜を堆積すること、および前記シリコン酸化膜を平坦にすることを含み、
前記伝導性材料および下層の前記制御サイドウォールスペーサゲートが共に制御ゲートを形成する請求項２２の製造方法。
【請求項４１】　前記第3のポリシリコン膜が約９０～１８０nmの厚さを有する請求項２２の製造方法。
【請求項４２】　前記ワード線をシリサイド化することをさらに含む請求項２２の製造方法。
【請求項４３】　前記ワード線をシリサイド化することをさらに含む請求項２２の製造方法。
【請求項４４】　半導体基板の表面上に窒化物含有膜を形成すること、
前記窒化物含有膜上を覆うように第１のポリシリコン膜を堆積すること、
前記第１ポリシリコン膜上を覆うように第2の窒化膜を堆積すること、
ワードゲートが、その間に間隙が残るように形成されるように、前記第１のポリシリコン膜および前記第2の窒化膜をパターニングすること、
前記ワードゲートのサイドウォール上に第１の絶縁膜を形成すること、
前記ワードゲートおよび前記ゲートシリコン酸化膜を覆うようにスペーサ膜を堆積すること、
除去可能なスペーサが前記ワードゲートのサイドウォール上に残るように、前記スペーサ膜を異方エッチングにより除去すること、
ビット拡散領域を形成するために、前記除去可能なスペーサをイオン打ち込みマスクとして機能させて、前記半導体基板内にイオンを打ち込むこと、
その後に前記除去可能なスペーサを取り除くこと、
前記ワードゲート上を覆い、前記間隙を充填する第2のポリシリコン膜を堆積すること、
前記第2のポリシリコン膜を前記第2の窒化膜の表面下までへこませること、
前記へこまされた第2のポリシリコン膜をシリサイド化し、そのシリサイド化され、へこまされた第2のポリシリコン膜が制御ゲートを形成すること、
前記シリサイド化され、へこまされた第2のポリシリコン膜上に酸化膜を堆積すること、
その後に前記ワードゲート上の前記第2の窒化層を除去すること、およびＭＯＮＯＳメモリ素子の前記製造を完了させるために、前記ワードゲート下で接続されるワード線を形成するような第3のポリシリコン膜を、前記基板上に堆積することを含むＭＯＮＯＳメモリデバイスの製造方法。
【請求項４５】　前記窒化物含有膜を形成する段階が、前記半導体基板上に、約3．6～5．0nmの厚さにまで第１のシリコン酸化膜を成長させること、
前記第１のシリコン酸化膜上に、約2～5nmの厚さを有するシリコン窒化膜を堆積すること、および前記シリコン窒化膜上に、約4～8nmの厚さを有する第2のシリコン酸化膜を堆積することを含む請求項４４の製造方法。
【請求項４６】　前記シリコン窒化膜を堆積する段階の前に、前記第１のシリコン酸化膜を窒化することをさらに含む請求項４５の製造方法。

【請求項４７】　前記第１のポリシリコン膜がＣＶＤにより約１５０～２５０ｎｍの厚さにまで堆積される請求項４４の製造方法。
【請求項４８】　前記第１の窒化膜がＣＶＤにより約５０～１００ｎｍの厚さにまで堆積される請求項４４の製造方法。
【請求項４９】　前記第１の絶縁膜が、前記ワードゲートのサイドウォール上に約５～１０ｎｍの厚さである請求項４４の製造方法。
【請求項５０】　前記スペーサ膜が、ポリシリコン、プラズマ窒化膜、プラズマ酸化窒化膜、およびホウ素リンガラス（ＢＰＳＧ）を含むグループ内のいずれかによって構成され、約３０～５０ｎｍの厚さである請求項４４の製造方法。
【請求項５１】　前記除去可能なスペーサを除去する工程の前に、
前記除去可能なスペーサによって被覆されない前記第２のシリコン酸化膜をエッチングすること、
前記窒化膜上に約４～６ｎｍの厚さにまで第３のシリコン酸化膜を堆積すること、および前記制御ゲートおよび前記ビット拡散領域間の結合容量が低減されるように、前記第３のシリコン酸化膜を酸化することによって、前記窒化膜上に約２０ｎｍの厚さを有する酸化膜を形成することをさらに含む請求項４４の製造方法。
【請求項５２】　前記除去可能なスペーサを除去する段階が乾式の化学的異法性エッチングを含む請求項４４の製造方法。
【請求項５３】　前記第２のポリシリコン膜が約３０～５０ｎｍの厚さを有する請求項４４の製造方法。
【請求項５４】　前期第２の絶縁膜が、ＣＶＤによって約１０ｎｍの厚さにまで堆積されたシリコン酸化膜を含む請求項４４の製造方法。
【請求項５５】　前期第２の絶縁膜が、ＣＶＤによって約１０ｎｍの厚さにまで堆積されたシリコン酸化膜を含む請求項４４の製造方法。
【請求項５６】　前記第３のポリシリコン膜が約１５０～２００ｎｍの厚さを有する請求項４４の製造方法。
【請求項５７】　間隙がその間に残るように、半導体基板の表面上のゲートシリコン酸化膜上にワードゲートを提供すること、
前記ワードゲートのサイドウォール上に除去可能なスペーサを形成すること、
浅いドープ領域を形成するために、前記除去可能なスペーサをイオン打ち込みマスクとして機能させて、前記半導体基板内にイオンを打ち込むこと、
その後に前記除去可能なスペーサを取り除くこと、
窒化膜注入領域として機能するような窒化物含有膜を下層にそれぞれ有するサイドウォールポリシリコンゲートを前記サイドワードゲート上に形成すること、
ビット拡散領域を形成ために、前記制御サイドウォールポリシリコンゲートをイオン打ち込みマスクとして機能させて、前記半導体基板内にイオンを打ち込むこと、
前記サイドウォールゲート上に絶縁膜を形成すること、
前記２つのワードゲート間の間隙を第２のポリシリコン膜で充填すること、
前記第２のポリシリコン膜をへこませること、
前記へこまされた第２のポリシリコン膜をシリサイド化すること、
前記へこまされ、シリサイド化された第２のポリシリコン層が前記下層のサイドウォールポリシリコンゲートと共に制御ゲートを形成するような酸化膜で、前記へこまされ、シリサイド化された第２のポリシリコン膜を被覆すること、およびフラッシュメモリデバイスの製造を完了させるために、前記ワードゲート下で接続されるワード線を形成するような第３のポリシリコン膜を、前記基板上に堆積することを含むフラッシュメモリデバイスの製造方法。
【請求項５８】　前記第１のポリシリコン膜が約１５０～２５０ｎｍの厚さを有する請求項５７の製造方法。
【請求項５９】　前記スペーサ膜が、ポリシリコン、プラズマ窒化膜、プラズマ酸化窒化膜、およびホウ素リンガラス（ＢＰＳＧ）を含むグループ内のいずれかによって構成される請求項５７の製造方法。
【請求項６０】　前記窒化物含有膜が、酸化シリコン膜の第１層、窒化シリコン膜の第２層、および酸化シリコン膜の第３層を含む請求項５７の製造方法。
【請求項６１】　前記除去可能なスペーサの除去の後に、約２０～５０ｎｍの深さを有する段差を前記半導体基板内に形成するために、前記半導体基板内にエッチングすることをさらに含む請求項５７の製造方法。
【請求項６２】　前記段差の角を丸める工程をさらに含む請求項５７の製造方法。
【請求項６３】　前記段差の角の丸め工程が、約１０００～１１００℃で、約６０秒間の高速熱焼きなましを含む請求項６２の製造方法。
【請求項６４】　前記段差の角の丸め工程が、約９００℃、約２００～３００ｍトール圧の水素内での焼きなましを含む請求項６２の製造方法。
【請求項６５】　前記ワードゲート縁から前記ビット拡散領域の縁までに限定されるチャネル長が約３０～５０ｎｍであり、これによってバリスティック電子注入が発生する請求項５７の製造方法。
【請求項６６】　半導体基板表面上のワードゲートと、
前記ワードゲートのサイドウォール上で絶縁膜によって前記ワードゲートから絶縁されたサイドウォール制御ゲートと、
前記サイドウォール制御ゲート下のＯＮＯ膜内に形成され、電子メモリ蓄積が実行される窒化物領域と、
前記ワードゲートおよび他のメモリセル内のワードゲートを覆って、これらを相互に接続し、さらに、絶縁膜に

よって前記サイドウォール制御ゲートから絶縁されて、当該サイドウォール制御ゲートを覆うポリシリコンワード線と、
前記半導体基板内で前記サイドウォール制御ゲートのそれぞれに隣接するビット線拡散領域とを含むMONOSメモリセル。
【請求項67】 各サイドウォールゲートが、絶縁膜によって前記他のメモリセルのサイドウォール制御ゲートから絶縁された請求項66のMONOSメモリセル。
【請求項68】 各制御ゲートが、前記ビット拡散領域および前記サイドウォール制御ゲートを覆うポリシリコン膜を2つのワードゲート間に含み、前記窒化物領域が前記サイドウォール制御ゲート下のみに形成された請求項66のMONOSメモリセル。
【請求項69】 前記ワードゲート縁から前記ビット拡散領域の縁までに限定されるチャネル長が約30～50nmであり、これによってバリスティック電子注入が発生する請求項66のNOMOSメモリセル。
【請求項70】 前記窒化物領域の一方が選択窒化物領域であり、他方の窒化物領域が非選択窒化領域であって、前記選択窒化物領域に近いビット線拡散領域がビット拡散領域であり、前記非選択窒化物領域に近いビット線拡散領域がソース拡散領域であって、セルの読み出し動作が、
前記非選択窒化物領域をオーバーライドすること、
ワードゲート閾値電圧、オーバードライブ電圧、および前記ソース拡散領域上の電圧の和を前記ワードゲートに供給すること、
前記選択窒化物領域に隣接する前記制御ゲートに、選択窒化物領域からの読み出しを可能にするのに充分な電圧を供給すること、および前記ビット拡散領域上の電圧準位を測定することによって前記セルを読み出すことによって実行される請求項66のMONOSメモリセル。
【請求項71】 前記メモリセルがMONOSメモリアレイ内の多数のメモリセルの1つであって、読み出される以外の全てのセルに0Vの制御ゲート電圧を供給することをさらに含む請求項70のMONOSメモリセル。
【請求項72】 前記メモリセルがMONOSメモリアレイ内の多数のメモリセルの1つであって、リークを防ぐために、読み出される以外の全てのセルに-0.7Vの制御ゲート電圧を供給することをさらに含む請求項70のMONOSメモリセル。
【請求項73】 前記ビット拡散領域上の電圧準位が、前記セルの複数の閾値準位の内の1つを代表する請求項66のMONOSメモリセル。
【請求項74】 前記窒化物領域の一方が選択窒化物領域であり、他方の窒化物領域が非選択窒化領域であって、前記選択窒化物領域に近いビット線拡散領域がビット拡散領域であり、前記非選択窒化物領域に近いビット線拡散領域がソース拡散領域であって、セルのプログラム動作が、
前記非選択窒化物領域をオーバーライドするために前記非選択制御ゲート上に高圧を供給すること、
前記選択窒化物領域の制御ゲート電圧を高めること、
前記ビット拡散領域上に一定の電圧を供給すること、
前記ワードゲート閾値電圧より大きな電圧を前記ワード線上に供給すること、およびその時にチャネル領域から前記選択窒化物領域への電子のバリスティック注入が生じるような、前記ソース拡散領域から前記ビット拡散領域への電流が流れるように、前記ソース拡散領域の電圧を低くすることによって実行される請求項66のMONOSメモリセル。
【請求項75】 前記ビット拡散線上の電圧を変えることによって複数の閾値がプログラムされる請求項74のMONOSメモリセル。
【請求項76】 前記メモリセルがMONOSメモリアレイ内の多数のメモリセルの1つであって、1つのワード線を共有する隣接したセルの窒化物領域を、そのセルに0Vの制御ゲート電圧を供給することによって無効にすることをさらに含む請求項74のMONOSメモリセル。
【請求項77】 前記制御ゲートの一方が選択制御ゲートであり、その下層の窒化物領域が選択窒化物領域であり、かつ、他方の制御ゲートが非選択制御ゲートであり、その下層の窒化物領域が非選択窒化物領域であって、前記選択窒化物領域に近いビット線拡散領域がビット拡散領域であり、前記非選択窒化物領域に近いビット線拡散領域がソース拡散領域であって、セルのプログラム動作が、
前記非選択窒化物領域をオーバーライドするように前記非選択制御ゲート上に高電圧を供給すること、および前記選択制御ゲート上の電圧を変えることによって実行される請求項66のMONOSメモリセル。
【請求項78】 前記メモリセルが、1つのワード線を共有するフラッシュメモリアレイ内の多数のセルの1つであり、前記制御ゲートあるいは前記ビット拡散領域のどちらかの電圧を変えることによって、複数のセルを異なる閾値で同時にプログラムすることをさらに含む請求項66のMONOSメモリセル。
【請求項79】 窒化物領域の1つのブロックの消去動作が、
前記ビット線拡散領域に正の電圧を供給すること、および前記ビット線拡散領域上の制御ゲートに負の電圧を供給することによって実行される請求項66のMONOSメモリセル。
【請求項80】 窒化物領域の1つのブロックの消去動作が、
前記半導体基板および前記ビット線拡散領域に負の電圧を供給すること、および前記制御ゲートに正の電圧を供給することによって実行される請求項66のMONOS

メモリセル。
【請求項８１】　半導体基板表面上のワードゲートと、
前記ワードゲートのサイドウォール上で絶縁膜によって前記ワードゲートから絶縁されたサイドウォール制御ゲートと、
前記サイドウォール制御ゲート下のＯＮＯ膜内に形成され、電子メモリ蓄積が実行される窒化物領域と、
前記ワードゲートおよび他のメモリセル内のワードゲートを覆って、これらを相互に接続し、さらに、絶縁膜によって前記サイドウォール制御ゲートから絶縁されて、当該サイドウォール制御ゲートを覆うポリシリコンワード線と、
前記半導体基板内で前記サイドウォール制御ゲートのそれぞれに隣接するビット線拡散領域とを含み、
前記窒化物領域の一方が選択窒化物領域であり、他方の窒化物領域が非選択窒化領域であって、前記選択窒化物領域に近いビット線拡散領域がビット拡散領域であり、
前記非選択窒化物領域に近いビット線拡散領域がソース拡散領域であって、セルの読み出し動作が、
前記非選択窒化物領域をオーバーライドすること、
ワードゲート閾値電圧、オーバードライブ電圧、および前記ソース拡散領域上の電圧の和を前記ワードゲートに供給すること、
前記選択窒化物領域に隣接する前記制御ゲートに、選択窒化物領域からの読み出しを可能にするのに充分な電圧を供給すること、および前記ビット拡散領域上の電圧準位を測定することによって前記セルを読み出すことによって実行されるＭＯＮＯＳメモリセルの書込み方法。
【請求項８２】　半導体基板表面上のワードゲートと、
前記ワードゲートのサイドウォール上で絶縁膜によって前記ワードゲートから絶縁されたサイドウォール制御ゲートと、
前記サイドウォール制御ゲート下のＯＮＯ膜内に形成され、電子メモリ蓄積が実行される窒化物領域と、
前記ワードゲートおよび他のメモリセル内のワードゲートを覆って、これらを相互に接続し、さらに、絶縁膜によって前記サイドウォール制御ゲートから絶縁されて、当該サイドウォール制御ゲートを覆うポリシリコンワード線と、
前記半導体基板内で前記サイドウォール制御ゲートのそれぞれに隣接するビット線拡散領域と含み、
前記窒化物領域の一方が選択窒化物領域であり、他方の窒化物領域が非選択窒化領域であって、前記選択窒化物領域に近いビット線拡散領域がビット拡散領域であり、
前記非選択窒化物領域に近いビット線拡散領域がソース拡散領域であって、セルの読み出し動作が、
前記非選択窒化物領域をオーバーライドすること、
ワードゲート閾値電圧、オーバードライブ電圧、および前記ソース拡散領域上の電圧の和を前記ワードゲートに供給すること、
前記選択窒化物領域に隣接する前記制御ゲートに、選択窒化物領域からの読み出しを可能にするのに充分な電圧を供給すること、および前記選択制御ゲート上の電圧を変える段階を含むＭＯＮＯＳメモリセルのプログラム方法。
【請求項８３】　半導体基板表面上のワードゲートと、
前記ワードゲートのサイドウォール上で絶縁膜によって前記ワードゲートから絶縁されたサイドウォール制御ゲートと、
前記サイドウォール制御ゲート下のＯＮＯ膜内に形成され、電子メモリ蓄積が実行される窒化物領域と、
前記ワードゲートおよび他のメモリセル内のワードゲートを覆って、これらを相互に接続し、さらに、絶縁膜によって前記サイドウォール制御ゲートから絶縁されて、当該サイドウォール制御ゲートを覆うポリシリコンワード線と、
前記半導体基板内で前記サイドウォール制御ゲートのそれぞれに隣接するビット線拡散領域とを含み、
窒化物領域の１つのブロックの消去方法が、
前記ビット線拡散領域に正の電圧を供給する段階、および前記ビット線拡散領域上の制御ゲートに負の電圧を供給する段階を含むＭＯＮＯＳメモリセルの消去方法。
【請求項８４】　半導体基板表面上のワードゲートと、
前記ワードゲートのサイドウォール上で絶縁膜によって前記ワードゲートから絶縁されたサイドウォール制御ゲートと、
前記２つのサイドウォール制御ゲート間の半導体基板内のビット線拡散領域と、
前記サイドウォール制御ゲート下の窒化物充填領域とを含むフラッシュメモリデバイス。
【請求項８５】　前記サイドウォール制御ゲート上の絶縁膜と、
前記制御ゲート上にあって、前記ワードゲートを接続するワード線とをさらに含む請求項８４のデバイス。
【請求項８６】　前記ワードゲートの縁から前記ビット拡散領域の縁までに限定される前記チャネル長が約３０～５０ｎｍであり、そこでバリスティック電子注入が生じる請求項８４のＭＯＮＯＳメモリセル。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、高集積される金属ポリシリコン－酸化膜－窒化膜－酸化シリコン膜（ＭＯＮＯＳ）メモリアレイの製造方法および高集積ＭＯＮＯＳメモリアレイに関する。
【０００２】
【従来の技術】不揮発性メモリには、浮遊ゲートおよびＭＯＮＯＳという２つの様式がある。従来の浮遊ゲート構造では、Ｆ－Ｎトンネリングあるいはソースサイド注入のどちらかによって、浮遊ゲート上に電子が格納される。従来のＭＯＮＯＳデバイスでは通常、メモリワード

ゲート下の酸化膜・窒化膜・酸化膜（ＯＮＯ）層内の直接トンネリングによって電子を格納する。電子はＯＮＯ積層膜の窒化膜に捕獲される。ＭＯＮＯＳトランジスタは、浮遊ゲートデバイスよりも１つ少ないポリシリコン膜しか必要としないので、プロセスが簡略化され、より一層密なアレイを得ることができる。

【０００３】

【発明が解決しようとする課題】ＭＯＮＯＳ構造は、一般的には、その中のＯＮＯ積層膜がワードゲートの下に堆積されるプレーナーデバイスである。プログラム動作用の直接トンネリングを利用するためには、ＯＮＯ膜の底部酸化膜の厚さは３．６ｎｍより薄くなければならない。しかしながら１９９８年に、クォータン　チャン（Kuo-Tung Chang）らによる「プログラミングのためにソースサイド注入を用いる新ＳＯＮＯＳメモリ（A New SONOS Memory Using Source Side Injection for Programming）」（IEEE Electron Letters、１９９８年７月、Vol.19,No.7）で、厚さ５．０ｎｍの底部酸化膜、サイドウォールポリシリコンゲート、およびソースサイド注入プログラムを有するＭＯＮＯＳ構造が初めて報告された。当該構造では、図１に示されるように、典型的なサイドウォールプロセスによって、ワードゲートの一方上にサイドウォールスペーサ２０が形成されて、従来のＭＯＮＯＳメモリセルに関するワードゲート下の代わりに、サイドウォールゲート下にＯＮＯ積層膜２２がある。ＳＯＮＯＳサイドウォール制御ゲート下のチャネル長は１００ｎｍより大きいので、プログラム機構は、より厚い底部酸化膜にもかかわらず、電子トンネリングよりも高速で低い電圧を必要とするようなソースサイド注入である。ソースサイド注入の間、サイドウォールゲートと選択／ワードゲートとの間隙にチャネル電位が形成される。チャネル電子３０は、前記間隙内で加速されて、ＯＮＯ膜内に注入するのに充分な熱電子になる。したがって、クォータン　チャンのＳＯＮＯＳメモリは、従来の直接トンネリングＭＯＮＯＳセルよりもすぐれたプログラム性能を達成することができる。

【０００４】ＳＯＮＯＳメモリセルは、そのスプリットゲート構造およびソースサイド注入プログラムの点でＭＯＮＯＳメモリの内で独特だけれども、その構造およびプログラムの主要部分は、従来のスプリットゲート浮遊ゲートデバイス用のそれらと類似である。どちらのセル様式も、ワードゲートとサイドウォールスペーサゲートとを並べて有する。サイドウォールゲートの利用と電子蓄積領域の構造に、非常に大きな違いがある。スプリットゲート浮遊ゲートセルでは、サイドウォールスペーサは、その上に電子が格納されるような浮遊ゲートである。ワードゲート、拡散領域、および浮遊ゲート間を接続する電気容量によって浮遊ゲート電圧が決定される。ＳＯＮＯＳセルでは、制御ゲートと呼ばれるサイドウォールスペーサ下の窒化物領域内に電子が格納される。窒化物領域の電圧は、前記サイドウォールゲート電圧によって直接的に制御される。

【０００５】１９９９年５月１７日に出願された、同一発明者の米国特許出願第０９／３１３，３０２号で、より高速なプログラム、およびより高い集積度を有する浮遊ゲートメモリセルが紹介された。図３Ａは前記高速プログラム、デュアルビット、高集積度のメモリセルの配列図であり、図３Ｂはその配置断面図である。このメモリ構造では、２つのサイドウォール浮遊ゲートを１つのワードゲートに組み合わせること（例えば、浮遊ゲート３１２、３１３とワードゲート３４１）、および互換性のあるソースドレイン拡散領域（３２１、３２２）をセル間に共有することによって高集積度が実現される。すなわち、１つのメモリセルは２つの電子記憶領域を有する。追加したポリシリコン線の「制御ゲート」は、拡散領域に対して平行に、かつワードゲートに対して直角に進む。制御ゲート（３３１、３３２）は浮遊ゲートに結合されて、一対の浮遊ゲートから個別に１つの浮遊ゲートを選択するように異なる制御方向を提供する。さらに、このメモリはバリスティック注入による高速プログラミングによって特徴づけられる。同じデバイス構造を用いて、サイドウォールゲートチャネルが適当な不純物断面（profile）を含み、４０ｎｍより短くされる場合には、注入機構が、ソースサイド注入から、バリスティック注入と呼ばれるような新しく、極めて一層有効な注入機構に変化する。エス．オグラ（S.Ogura）による、１９９８年発行のＩＥＤＭ、９８７頁の「ＥＥＰＲＯＭ／フラッシュのためのバリスティック直接注入を有するステップ・スプリット・ゲートセル」（Step Split Gate Cell with Ballistic Direction Injection for EEPROM/Flash）で、バリスティック注入機構が証明された。図２Ａでは、浮遊ゲートメモリセルの、バリスティック注入（線２５）および従来のソースサイド注入（線２７）の結果が比較される。それらの構造は非常に似ているけれども、制御ゲートが１００ｎｍの場合、注入機構はソースサイド注入である。しかしながら、図２Ｂに示されるように、バリスティック注入（線３５）用に必要な短いチャネル長を満たすように、チャネル長が４０ｎｍにまで低減されると、プログラム速度が、同じバイアス条件下で３倍になるか、またはソースサイド注入（線３７）用に必要な浮遊ゲート電圧の半分で加速される。

【０００６】対照的に、クォ　タン　チャンのＳＯＮＯＳメモリ構造のサイドウォールチャネル長は２００ｎｍであり、したがってプログラム機構はソースサイド注入である。すなわち、短いチャネル長と注入機構との間には重大な相関関係が存在する。

【０００７】

【発明を解決するための手段】発明の概要

本発明では、２つまたは３つのポリシリコンスプリットゲート・サイドウォールプロセスによって、高速低電圧

バリスティックプログラム、超短チャネル、超高集積度、デュアルビット多準位のフラッシュメモリが達成される。3～5Vの低いプログラム電圧で高い電子注入効果および極めて高速なプログラムを提供するようなバリスティック注入（エス．オグラ：S. Ogura）を伴なう、40nmより短い超短制御ゲートチャネルを有するツインMONOSセル構造によって、本発明の構造および動作が実現可能である。セル構造は、（ｉ）ワードゲートの両サイド上の酸化膜－窒化膜－酸化膜（ONO）の積層膜上にサイドウォール制御ゲートを配設すること、および（ｉｉ）自己整合によって制御ゲートおよびビット拡散領域を形成し、高集積度用メモリセル間で制御ゲートおよびビット拡散領域を共有することによって実現される。本プロセスに用いられる主要素は、

（ｉ）超短チャネルを製造するための除去可能なサイドウォールプロセスおよび段差付きあるいは無しのサイドウォール制御ゲート

（ｉｉ）蓄積窒化膜上の制御ゲートと自己整合された拡散領域、および制御ゲートと同じ方向に延設されたビット線拡散領域である。

【0008】本発明の高速プログラム、低電圧、超高集積度、デュアルビット、多準位のMONOS NVRAMの特徴は、

1．制御ゲート下のONO膜内の窒化物領域内の電子メモリ

2．セル毎に2つの窒化膜メモリ要素がある高集積デュアルビットセル

3．高集積デュアルビットセルが各窒化物領域内に多準位を記憶できること

4．ワードゲートおよび制御ゲートによって制御される低電流プログラム

5．制御可能な超短チャネルMONOSを利用するバリスティック注入による高速、低電圧プログラム

6．選択されていない隣接の窒化物領域およびメモリセルのメモリ蓄積状態の影響をマスキングアウトする間に、多準位をプログラムし、かつ読み出すためのサイドウォール制御ポリゲートを含む。

【0009】バリスティックMONOSメモリセルは、次のような配列で整列される。各メモリセルは1つのワードゲート用の2つの窒化物領域、および2分の1のソース拡散領域、2分の1のビット拡散領域を含む。制御ゲートは、離して区切られるか、または同じ拡散領域上で共有される。拡散領域はセル間で共有されて、サイドウォール制御ゲートに対して平行に、かつワード線に対して垂直に延設される。

【0010】多準位記憶のための作動条件の概要が図3Bに示される。読み出し中は次の条件が満たされなければならない。選択されたメモリセル内の選択されない制御ゲートの電圧は、制御の閾値電圧とソース電圧との和より大きくなければならない。一対の制御ゲート内のワード選択ゲートは、ワードゲートの閾値電圧と0．5V付近のオーバーライド増加分（override delta）とソース電圧との和（Vt-wl＋Voverdrive＋Vs）まで高められる。関連する制御ゲートを0Vまで低減することによって、選択されないMONOSセルが無効にされるであろう。プログラム条件は、ワード線電圧が、その閾値と、低電流プログラムのためのオーバードライブ電圧増加分との和より大きいこと、選択された一対の制御ゲートのいずれもがVt-high（多準位閾値の範囲内での最高閾値電圧）とオーバーライド増加分との和より大きいこと、および同一ワード線電圧を共有する隣接メモリセルが制御ゲートのみを調整されることによって無効とされることである。

【0011】

【発明の実施の形態】好ましい実施例の説明本発明によって、2つの窒化膜メモリ要素および2分割された制御ゲートを有するバリスティックツインMONOSメモリセルのための製造方法が提供される。その方法は、フラットチャネルを有するデバイス、および／あるいはMONOSセル内の窒化膜の下にステップチャネルを有するデバイスに適用できる。

【0012】

【実施例】浅溝分離、pウェル、およびnウェルの形成手順は、従来のCMOS手法のそれと同じなので示されない。ポリシリコンワードゲートもまた、図4Aに示されるような従来のCMOSプロセスによって形成される。ワードゲートを形成するために、メモリゲート酸化シリコン膜221が、約5～10ナノメートルの間の厚さに形成される。それから、ゲート材料用に、化学気相成長法（CVD）によって、約150～250nmの厚さのポリシリコン245が堆積される。化学的機械研磨（chemical mechanical polishing：CMP）に対するエッチングストッパとして後に使用される窒化膜232が、CVDよって約50～100nmの厚さに堆積される。標準的なCMOSプロセスでメモリワードゲートを形成する。すなわち、フォトレジストプロセス、露光および現像を伴うマスキングプロセス、および反応イオンエッチング（RIE）による窒化膜232およびポリシリコン245への水平エッチングが実現される。ホウ素202が追加的に、浮遊ゲート下のVTを調整するために、1平方センチメートルあたり3E12～3E13のドーズ量で、かつ低エネルギ（約10KeVエネルギより低い）で注入される。ワードゲートを形成するために用いられたフォトレジストの除去後に、ワードゲートは図4Aに示されるようになる。

【0013】図4Bに示されるように、サイドウォールポリシリコンの表面に、約5～10nmの薄い酸化シリコン膜234が熱的に成長されるか、または二酸化シリコンおよび／あるいは窒化シリコンフィルムが共通のCVD法によって堆積される。それから、制御可能な短チ

ャネルを限定し、高い電子注入効果による高速プログラミングを提供する除去可能なサイドウォールの製造プロセスが実行される。典型的には３０～５０ｎｍの厚さを有する１つの薄いポリシリコン膜が堆積される。さらに、図４Ｂに示されるように、ワードゲート２４５の両サイド上に除去可能なサイドウォールスペーサ２４２を形成するような、垂直または異方性のポリシリコンエッチングが施される。ヒ素などのＮドーパント２０３の注入が、１０～１５ＫｅＶで３Ｅ１３～４Ｅ１３／ｃｍ^{2}のイオンドーズ量で実行される。つまり、ポリシリコン膜の厚さが制御ゲート下の有効チャネル長を決定する。

【００１４】図４Ｃに関しては、乾式の化学的異方性エッチングによって、除去可能なサイドウォールスペーサ２４２が徐々に除去される。この段階での典型的なエッチング環境は、ＨＢｒ／Ｃｌ$_2$／Ｏ$_2$である。次に、（例えば水酸化アンモニウム水で）緩衝されたフッ化水素酸（ＢＨＦ）、気相ＨＦ、あるいはＣＦ$_2$／Ｏ$_2$のような反応イオンエッチングによって、底部酸化シリコン膜２２１が徐々に食刻（etching out）される。酸化膜（Ｏ）－窒化膜（Ｎ）－酸化膜（Ｏ）の積層膜２３０が形成される。膜２３０は、簡略にするために３層では示されない。底部酸化膜は熱成長され、その厚さは直接トンネリングの限界値（３．６ｎｍ）より僅かに厚い３．６～５ｎｍであり、ＣＶＤによって堆積された窒化シリコン膜は約２～５ｎｍであり、さらに頂部酸化膜はＣＶＤ堆積法によって堆積されて、それは約４～８ｎｍである。頂部酸化膜の品質を高めるために、熱酸化を加えることができる。また底部酸化膜の信頼性を高めるために、窒化膜の堆積の前にＮ$_2$Ｏ環境内での短時間の窒化を加えることもできる。

【００１５】ここで、約３０～５０ｎｍのリン添加ポリシリコン薄膜および６０～１００ｎｍのタングステンシリサイドがＣＶＤにより堆積される。ポリシリコンおよびタングステンシリサイドの積層膜は制御サイドウォールスペーサゲートになる。図４Ｃに示されるように、サイドウォール制御ゲート２４０を形成するために、垂直、異方向性反応エッチングが実行される。酸化膜－窒化膜－酸化膜の積層膜も貫通してエッチングされて、サイドウォール制御ゲート下のみに、このＯＮＯ膜２３０が残る。

【００１６】厚さ約１０ｎｍの酸化シリコン膜あるいは窒化膜の薄いＣＶＤ膜２３３が堆積される。図４Ｃに示されるように、n$^+$注入領域２０４のためのリンおよび／あるいはヒ素が、３Ｅ１４～５Ｅ１５イオン／ｃｍ^{2}のドーズ量で注入される。合計の厚さは９０～１５０ｎｍであり、それは有効な制御ゲートチャネル長とn$^+$注入領域の外部拡散領域の和に等しい。

【００１７】変形例としては、サイドウォールスペーサゲート２４０は、ポリシリコンとタングステンシリサイドの積層膜ではなく、単純に、リンあるいはヒ素添加ポリシリコン膜であってよい。制御ゲートをシリサイド化して低抵抗化するのであれば、図４Ｄに示されるように、n$^+$注入領域の形成、および厚さ約１０ｎｍの酸化シリコン膜あるいは窒化膜の薄いＣＶＤ膜２３３の堆積の後に、ゲート２４０上にサイドウォール酸化膜スペーサ２３３を形成するために垂直反応イオンエッチングが実行される。典型的なシリサイド化では、プラズマスパッタ法（sputtering）によって約１０ｎｍのコバルトあるいはチタンが堆積され、約６５０℃で高速熱焼きなまし（アニール）が実行される。ゲート２４０および拡散領域２０４の頂部上のシリサイド層２４１の構成が図４Ｄに示される。図４Ｄにはシリサイド層２４１が図示されるが、それは必須ではない。動作、読み出し、プログラム、および消去の全モードのパフォーマンスを向上するために、制御ゲート線あるいは拡散領域線のＲＣ時定数を低減することは一つの選択である。

【００１８】混成障壁用の酸化膜および／あるいは窒化膜２３５がＣＶＤによって堆積される。次に、間隙を埋めるためにＣＶＤ酸化シリコン膜あるいはＢＳＧの膜２４７が堆積される。間隙充填材はＣＭＰによって磨かれて窒化膜２３２になる。

【００１９】変形例としては、間隙充填材２４７は、サイドウォールゲートのＲＣ時定数あるいは必要に応じてはビット拡散領域を低減するために用いることができるような、ポリシリコンあるいはタングステンなどの導電性材料であることができる。導電膜がＣＭＰによって磨かれて窒化膜２３２になる時に、導電膜は垂直反応イオンエッチングによって数百ナノメートル（５０ｎｍ）へこまされる。次に、ＣＶＤにより二酸化シリコン膜（約５０ｎｍ）が堆積されて、図４Ｅの２３６によって示されるようにＣＭＰが実行される。

【００２０】図４Ｅの窒化膜２３２は、H_3PO_4によって、あるいは乾式の化学的エッチングによって選択的にエッチングされる。１５０～２００ｎｍの厚さのポリシリコン膜がＣＶＤによって堆積される。当該ポリシリコン膜２４８および下層のポリシリコンワードゲート２４５が、通常のフォトレジストおよびＲＩＥプロセスによって限定される。この時点の構造が図４Ｆに示される。

【００２１】隣接するワード線ゲートを結合することによって、ポリシリコン膜２４８がワード線ワイヤとして機能する。最終的なメモリセルがこの時点で完成される。シート抵抗を低減するために、当該ワードポリシリコン膜はチタンあるいはコバルトでシリサイド化され得る。メモリセルの典型的な平面図が図４Ｇに示される。浅溝分離領域が図４Ｇに領域２０９で示される。

【００２２】前述のプロセスは、非常に短いチャネル（３０～５０ｎｍ）を有するプレーナーチャネル浮遊ゲートの製造を説明する。少しのプロセス段階を変更および追加することによって、プレーナー構造と同じ集積配列を用いて、より効果的なバリスティック注入を伴なう

ステップスプリット構造が構成され得る。本発明の、この第2の実施例は図5B、5C、および5Fを参照して詳述される。

【0023】ドープされたポリシリコンを垂直にエッチングすることによって除去可能なサイドウォールスペーサ242を形成した後に、図4Bに対応するように、酸化シリコン膜221が垂直にエッチングされる。ステップスプリットメモリセルを形成するためのプロセス変更は、エッチングをシリコン基板内におよそ20～50nmの深さまで続けることから始まる。次に、図5Bに示されるようにポリサイドウォールをマスクとして用いて、10～15KeVのエネルギでドーズ量が3E13～4E13/cm²であるようなN領域203を形成するために、段差部の底部にヒ素が僅かに埋め込まれる。次に、除去可能なN⁺添加ポリシリコンスペーサが、湿式エッチング(HNO_3/HF/Acitic酸、あるいはH_3PO_4またはNH_4OH)か乾式プラズマエッチングのどちらかによって、ドープされたバルクN⁻領域まで選択的に除去される。この除去可能なスペーサエッチング中のバルクエッチングは、段差状エッチングの一部として含まれ得る。除去可能なポリシリコンスペーサ下に残されたゲート酸化膜221を徐々に食刻した後に、シリコン表面が一掃される。シリコン内への総段差は約20～50nmでなければならない。段差部の角が尖っている場合には、約60秒間の約1000～1100℃での高速熱焼きなまし(RTA)による角の丸めが選択的に加えられるか、あるいは900℃、200～300mトール圧での水素焼きなましが実行され得る。これらの変更または追加の後に、製造工程は前述の手順に戻る。

【0024】図5Cに示したように、酸化膜-窒化膜-酸化膜の積層膜が形成される。膜230は簡明にするために3層では示されない。底部酸化膜は熱酸化により形成され、その厚さは直接トンネリングの限界(3.6nm)よりも僅かに厚い3.6～5nmである。CVDによって堆積された窒化シリコン膜は約2～5nmである。さらに頂部酸化膜がCVDによって堆積されて、それは約4～8nmである。頂部酸化膜の品質を高めるために熱酸化を加えることができる。また底部酸化膜の信頼性を高めるために、窒化膜を堆積する前に酸化窒素環境内での短時間の窒化を加えることもできる。

【0025】次に、制御ゲートになるリン添加ポリシリコン膜が90～180nmの厚さで堆積されて、図5Cに示されるように、サイドウォールゲート240を形成するために、垂直あるいは異方性のエッチングが実行される。プレーナースプリットデバイス用に与えられた製造工程を続けることによって、図5Fに示されるように、ステップスプリットデバイスを製造できる。当該サイドウォールポリシリコンゲートは、シリサイド化されるか、または平坦チャネルMONOSツインセルの第1の実施例で実現されるような耐熱性シリサイドによって置き換えることができる。

【0026】プレーナーおよびステップデバイスの両方用の前述の製造工程においては、除去可能なサイドウォールスペーサ242は、ポリシリコンの代わりに、プラズマ窒化膜、酸化膜またはホウ素リンガラス(BPSG)でも良い。なぜならば、熱酸化シリコン膜に対する、H_3PO_4酸または希釈HF内でのエッチング割合は非常に高い(例えば少なくとも10-100倍)からである。

【0027】本発明の第3の実施例が図6A-6D、および6Fを参照して述べられる。本発明の第3の実施例では、2つのサイドウォールスペーサの代わりに単一の大きなスペーサを用いることによって制御可能性が喪失され、その結果、僅かにプログラム速度が遅くなるものの、第1の実施例のプレーナーツインMONOSメモリセルの製造工程が簡略化される。通常のCMOSプロセスからの変更がワードゲートポリシリコン245の堆積の前から始まる。図6Aの酸化膜-窒化膜-酸化膜(ONO)の積層膜230が形成される。膜230はここでも簡略化のために3層では示されない。底部酸化シリコン膜は約3.6～5nmの厚さで熱酸化により形成されるのが好ましく、CVDによって堆積された窒化シリコン膜は約2～5nmであり、頂部の酸化膜はCVDによって堆積されて、約5～8nmの厚さである。ポリシリコンおよび除去可能なサイドウォールスペーサが連続してエッチングされないように、頂部のCVD酸化膜は第1および第2のプロセス実施例に較べて僅かに厚い。次に、CVDによってゲート材料用のポリシリコン245が堆積され、引き続きCVD窒化シリコン膜232が約50～100nmの厚さに堆積される。

【0028】次に、メモリゲート245を形成するために、フォトレジスト膜が形成され、露光および現像を伴なうマスキングプロセスが実行される。次いで、下層の積層膜230内の頂部の酸化シリコン膜をエッチングストッパとして、反応イオンエッチング(RIE)によってポリシリコン膜が垂直にエッチングされる。次に、図6Aに示されるように、ホウ素202が低エネルギ(10KeVより低い)、かつ5E12～2E13イオン/cm²のドーズ量で追加的にイオン打ち込みされ、ヒ素もまた、前記ホウ素と同じ程度の約5E12～1.5E13KeVで同時に浅く打ち込まれる。ヒ素の影響により、チャネル閾値が非常に低いけれども、短チャネル領域内にチャネル電位降下を生じるための不純物は多く存在する。

【0029】約5nmのシリコン薄膜234がポリシリコンの側面上に熱酸化で形成されるか、あるいは同様にCVDにより堆積される。次に、典型的には約90～150nmの厚さを有する除去可能なポリシリコン膜が堆積される。さらに、図6Bの除去可能なサイドウォール

スペーサ２４３を形成するような、垂直あるいは異方向性のポリシリコンエッチングが実行される。このスペーサは第１および第２の実施例のスペーサより厚い。次に、N^+注入領域２０４を形成するために、酸化膜－窒化膜の積層膜を貫通して、１Ｅ１５～５Ｅ１５／ｃｍ^{2}のドーズ量、かつ２０～５０ＫｅＶのエネルギレンジで、ヒ素イオンが打ち込まれる。低電力での高いバリスティック注入効果のために、焼きなましの温度と時間（８５０～９００℃で５～２０秒）で外部拡散領域を調整することによって、ワードゲートのエッジからN^+注入領域２０４のエッジまでで定義されるチャネル長が、約３０～５０ｎｍ（電子の平均自由工程の３～４倍）に設計される。

【００３０】その後、乾式の化学的等方性エッチングによって、除去可能なサイドウォールスペーサ２４３が徐々に除去される。この段階での典型的なエッチング環境は$HBr/CL_2/O_2$である。緩衝フッ化水素酸によって、窒化膜上の露出された酸化シリコン膜が徐々に食刻される。図６Ｃに示される積層膜ＯＮＯ２３０内の頂部酸化膜に代わって、ＣＶＤによって約４～６ｎｍの新しい酸化シリコン膜２４４が堆積される。頂部酸化膜の品質を高めるために、頂部膜が堆積された後に熱酸化が加えられる。

【００３１】変形例としては、除去可能なサイドウォールスペーサ２３４の除去の前に、ＲＩＥによって酸化膜－窒化膜の露光された頂部２層がエッチングされる。次に、頂部酸化膜を改質するために、ＣＶＤおよび連続的な熱酸化によって約４～６ｎｍの新しい酸化膜が堆積される。ウェットな二酸化環境内での約８５９～９００℃で２０分の前記酸化プロセス中に、図６Ｄに２４４で示されるように、n^+注入領域上の窒化膜除去領域上に約２０ｎｍの酸化膜が追加的に形成される。この厚い酸化膜が、制御ゲート２４０とビット拡散領域２０４との間の接続静電容量を低減する。

【００３２】ワードポリシリコン２４５と頂部窒化膜２３２の高さの和よりもわずかに厚い、およそ３００ｎｍのポリシリコン膜が堆積されて、エッチングストッパとして窒化膜を用いたＣＭＰが実行される。次に、充填されたポリシリコン膜２４０が、垂直、異方向性反応イオンエッチングによって、約５０ｎｍへこまされる。次に、約１０ｎｍの薄いチタンあるいはコバルトが堆積されてシリサイド化が実行される。シリサイド膜２４１は制御ゲート抵抗を低減するためのものである。２３６によって図示されるように、ＣＶＤによる二酸化シリコンの堆積およびＣＭＰが再度実行される。この時点でのデバイスの断面が図６Ｃおよび６Ｄに示される。

【００３３】次に、H_3PO_4あるいは乾式の化学的エッチングによって、窒化膜２３２が選択的にエッチングされる。約１５０～２００ｎｍの厚さを有するポリシリコン膜２４８がＣＶＤによって堆積される。通常のフォトレジストおよびＲＩＥプロセスによって、当該ポリシリコン膜および下層のワードゲートポリシリコン２４５が加工される。この時点での構造が図６Ｄに示される。

【００３４】隣接するワード線ゲートを結合することによって、ポリシリコン膜２４８がワード線ワイヤとして機能する。最終的なメモリセルがこの時点で完成される。シート抵抗を低減するために、当該ワードポリシリコン膜は、チタンあるいはコバルトでシリサイド化される。メモリセルの典型的な平面図が図４Ｇに示される。浅溝分離領域が領域２０９によって提供される。これらの臨界寸法が、臨界寸法が低減されるような技術で決定されることが理解される。

【００３５】前述の実施例においては、本発明のメモリ集積度を高めるために、２つの用法が組み合わされていた。第１の用法では、できるだけ多くのセル要素を共有することによって、集積度が２倍より大きい。１つのワード選択ゲートが２つの窒化膜蓄積領域間で共有され、制御ゲート線と同じソース線／ビット線が接合セル間で共有される。第２の用法では、複数の閾値が制御ゲート下の窒化物領域に記憶され、各閾値間のマージンを適正に保ちながら、高集積度アレイを可能にする多準位の感知およびプログラムを実現するために、所定の電圧および制御条件が開発されている。

【００３６】多準位記憶用の動作方法

以下に詳述される手順は、２ビット以上の多準位記憶ばかりでなく、制御ゲート下の窒化物領域内に記憶されるための、$V_{t\text{-}hi}$および$V_{t\text{-}low}$がそれぞれ閾値電圧の最高値および最低値であるような単一ビット／２準位記憶用法にも適用される。メモリセルのデュアルビット性は、単一のワードゲートに組み合わされた２つの窒化物領域の関連およびセル間のソースおよびドレイン領域の互換性に由来する。当該セル構造はサイドウォール堆積プロセスによって得られ、製造および動作の概念は、ステップスプリット・バリスティックトランジスタおよび／あるいはプレーナースプリットゲート・バリスティックトランジスタのどちらにも適用することができる。ステップスプリットおよびプレーナーバリスティックトランジスタは、低いプログラミング電力、高速プログラミング、および薄い酸化膜を有する。

【００３７】プレーナースプリットゲートバリスティックトランジスタアレイの断面が図７Ｂに示される。ワードゲート３４０、３４１、および３４２は全て第１準位ポリシリコン内に形成され、相互に接続されてワード線３５０を形成する。ワードゲート３４０、３４１、および３４２の両サイド上に堆積される一対のサイドウォールの下にＯＮＯが形成される。各サイドウォール下のＯＮＯ膜内の窒化膜は、電子メモリ用の事実上の領域である。これらの窒化物領域は、図７Ｂおよび７Ｃの３１０、３１１、３１２、３１３、３１４、３１５である。周辺復号化回路を単純にするために、プロセス実施例

３、および間隙充填材料２４７が導電体であるような実施例１および２によって単一制御ゲート３３０、３３１、３３２、３３３を形成することにより、同じ拡散領域を共有する２つのサイドウォール制御ゲートが結合される。１つの拡散領域を共有する２つのサイドウォールゲートがお互いに隔離される（間隙充填材料が絶縁材である）ようなプロセス実施例１および２の場合には、メモリアレイのワイヤアウトサイドでこれら２つのゲートを電気的に接続することが実現可能である。個別のサイドウォールゲートを制御ゲートとしてメモリアレイを操作することも可能だが、周辺論理回路は、高集積度メモリの利点を損なうような、さらなる負担となるであろう。

【００３８】窒化物領域３１１および３１２は制御ゲート３３１を共有し、窒化物領域３１３および３１４は制御ゲート３３２を共有する。メモリセル３０１は、ソース拡散領域３２１およびビット拡散領域３２２を有し、そのソース拡散領域とビット拡散領域との間に連続する３つのゲート、すなわち下層に窒化物領域３１２を具備する制御ゲート３３１、ワードゲート３４１、および下層に窒化物領域３１３を具備する他方の制御ゲート３３２を有するように説明することができる。ワードゲート３４１は単純論理オン／オフスイッチであり、制御ゲートは、読み出し中の選択された窒化物領域の電圧状態を個別に出力することを可能にする。同じワードゲートを共有する２つの窒化物蓄積領域は、本明細書において以下「窒化物蓄積領域ペア」と表現される。単一メモリセル３０１内では、窒化物蓄積領域ペア内の１つの窒化物蓄積領域３１３が、読み出しアクセスあるいはプログラム動作のために選択される。「選択された窒化物蓄積領域」３１３とは、選択された窒化膜ペアのうちの選択された窒化物領域のことである。「選択されない窒化物蓄積領域」３１２とは、選択された窒化物蓄積領域ペアのうちの選択されない窒化物蓄積領域のことである。「近位隣接窒化物蓄積領域」３１１および３１４とは、選択されたメモリセル３０１に最も隣接するような選択されないメモリセル内の窒化膜充填ペアの窒化物蓄積領域のことである。「遠位の選択されない隣接窒化物蓄積領域」３１０および３１５とは、同一の選択されない隣接メモリセル窒化物蓄積領域ペア内の隣の選択されない隣接窒化物蓄積領域の反対側の窒化物蓄積領域のことである。さらに、選択されたメモリセルの「ソース」拡散領域３２１は、選択された窒化物蓄積領域からの２つのメモリセル拡散領域であり、選択された窒化物蓄積領域に最も接近した接合部は、「ビット」拡散領域３２２と呼ばれる。

【００３９】本発明では、一組の窒化物蓄積領域から一方の窒化物蓄積領域の働きを消去するために、制御ゲート電圧が操作される。制御ゲート電圧の３つの状態、すなわち「オーバーライド（over-ride）」、「エクスプレス（express）」、および「抑止（suppress）」がある。制御ゲート電圧状態は、ワード線電圧の合計が２．０Ｖになり、「ビット」拡散領域電圧が０Ｖであり、かつ「ソース」拡散領域電圧の合計が１．２Ｖになるように説明される。与えられた電圧は、プロセス技術の特徴に基づく多数の適用可能例のうちの１例であり、いかなる限定でもないということが理解されなければならない。オーバーライド状態では、制御ゲート下のチャネルが窒化物領域内に蓄積された電荷に関わらず導電化されるように、Ｖ（ＣＧ）が高電圧（～５Ｖ）まで高められる。エクスプレス状態では、制御ゲート電圧が約Vt-hi（２．０Ｖ）まで高められ、制御ゲート下のチャネルは、窒化物領域のプログラム状態に依存して導電化されるであろう。抑止モードでは、下層のチャネルの導電化を抑止するために、制御ゲートが０Ｖに設定される。

【００４０】表１は、選択された窒化物領域３１３の読み出し中の電圧である。

【００４１】

【表１】

選択されたＦＧ＝３１３の読み出し用電圧

Vd0 320	Vcg 0 330	Vwl 340	Vd1 321	Vcg 1 331	Vwl 341	Vd2 322	Vcg 2 332	Vwl 342	Vd3 323	Vcg 3 333
0*	0	2.5	1.2	5	2.5	~0	2.5	2.5	0*	0

【００４２】閾値電圧がわずかに負の場合、わずかに負の制御ゲート電圧（約－０．７Ｖ）で窒化膜閾値領域を抑止することができる。

【００４３】図３Ｃに示される窒化物領域３１３の読み出し動作中、ソース線３２１は、ある中間の電圧（～１．２Ｖ）に設定されることができ、ビット線３２２はあらかじめ０Ｖに設定されてよい。さらに、選択された窒化物蓄積領域から読み出すためには、次の条件が満たされなければならない：１）ワード選択ゲート電圧が０Ｖから、ワード選択ゲートの閾値電圧（Vt-wl＝０．５Ｖ）とソース電圧（１．２Ｖ）との合計よりも大きな増加分である電圧（２．５Ｖ）まで高められなければならない、および２）選択された窒化物蓄積領域上の制御ゲートの電圧がVt-hi（「エクスプレス」）に近くなければならない。選択されない窒化物蓄積領域上の制御ゲートの電圧は、ソース電圧＋Vt-hi（「オーバーライド」）より大きくなければならない。選択されない隣接窒化物蓄積領域上の制御ゲートの電圧は、ゼロ（「抑止」）でなければならない。シリアルあるいはパラレル読み出しのそれぞれにおいて、窒化物蓄積領域３１３の

閾値電圧に対応するバイナリ値を決定するために、ビット拡散領域３２２の電圧がセンス増幅器によって監視され、切換え可能な基準電圧、あるいはそれぞれ異なる基準電圧を有する複数のセンス増幅器と比較される。つまり、選択されたメモリセル内の選択されない窒化物領域をオーバーライドし、次に隣接セルの選択されない窒化物領域を抑止することによって、個々の選択された窒化物領域の閾値状態が決定される。

【００４４】電子が酸化膜を貫通して窒化膜上に注入されるバリスティックチャネルのホットエレクトロン注入用に、電子が高いソースドレイン電位によって励起される。プログラムされた閾値電圧の大きさは、ソースドレイン電位およびプログラム時間によって制御される。表２は、選択された窒化物領域３１３に対して複数の閾値電圧をプログラムするための電圧を示す。これらの電圧は、単にプログラム方法の説明をするための例であって、いかなる限定でもない。表２Ａでは、窒化物蓄積領域３１２および３１３をオーバーライドするために、選択されたメモリセル３０１に関連する制御ゲート３３１、３３２が高電圧（５Ｖ）まで高められる。

【００４５】

【表２Ａ】

選択された窒化物蓄積領域３１３のビット拡散領域手法プログラム

Vt Data	Vd0 320	Vcg 0 330	Vwl 340	Vd1 321	Vcg 1 331	Vwl 341	Vd2 322	Vcg 2 332	Vwl 342	Vd3 323	Vcg 3 333
00	0	0	2.0	~0	5	2.0	5	5	2.0	0	0
01	0	0	2.0	~0	5	2.0	4.5	5	2.0	0	0
10	0	0	2.0	~0	5	2.0	4.0	5	2.0	0	0

【００４６】所望の閾値のプログラムはビット拡散領域３２２によって決定される。２．０Ｖ、１．６Ｖ、および１．２Ｖの閾値をプログラムするために、ビット拡散領域３２２がそれぞれ５Ｖ、４．５Ｖ、および４．０Ｖに固定される。ワード線３５０の電位がワードゲート３４１の閾値近くに高められた時に、高エネルギの電子がチャネル内に放出されて、注入が始まる。隣接するメモリセル内でのプログラムを防止するために、遠位の隣接制御ゲートは０Ｖに設定されるので、隣接メモリセルのチャネル内には電子が存在しないであろう。つまり、当該高集積度メモリアレイ用のビット拡散領域の電位制御によって、多準位閾値プログラムが達成される。例えば、１．２Ｖ、１．６Ｖ、および２．０Ｖをプログラムするために、それぞれ４．５Ｖ、５Ｖ、および５．５Ｖにワード線電圧を変更することによって複数の閾値をプログラムすることもできる。

【００４７】他の適用可能なプログラム方法は、異なる閾値を得るために制御ゲート電圧を変更するものである。多準位が制御ゲート電圧によって得られるものであるならば、選択されたメモリセル３０１内の選択されない制御ゲート３３１は、窒化物領域３１２をオーバーライドするために５Ｖにまで高めて設定されるであろう。閾値電位１．２Ｖ、１．６Ｖ、および２．０Ｖを得るために、選択された窒化物領域３１３上の制御ゲート３３２は、それぞれ４．５Ｖ、５Ｖ、および５．５Ｖに変更されるだろう。

【００４８】多準位プログラム用に説明された電圧条件に対する第４のプログラム方法が表２Ｂに示されており、選択された制御ゲート電圧はビット電圧に一致し、それぞれVd＝５Ｖ、４．５Ｖ、４．０ＶとVcg＝５Ｖ、４．５Ｖ、４．０Ｖである。

【００４９】

【表２Ｂ】

選択された窒化物蓄積領域３１３の制御ゲートービット方法プログラム

Vt Data	Vd0 320	Vcg 0 330	Vwl 340	Vd1 321	Vcg 1 331	Vwl 341	Vd2 322	Vcg 2 332	Vwl 342	Vd3 323	Vcg 3 333
00	0	0	2.0	~0	5	2.0	5	5	2.0	0	0
01	0	0	2.0	~0	4.5	2.0	4.5	4.5	2.0	0	0
10	0	0	2.0	~0	4.0	2.0	4.0	4.0	2.0	0	0

【００５０】プログラム電流が低いために、かつ前述のプログラミング構成によって、並列動作で同じワード線上の複数のセルをプログラムすることができる。さらに、ビット拡散領域あるいは制御ゲートのプログラム方法が用いられる場合には、周辺の復号化回路によって、複数の閾値が同時にプログラムされることもできる。しかしながら、適当な絶縁性を得るためには、選択されたメモリセルが、それらの間に２つ以上ものメモリセルを具備しなければならないことに注意しなければならない。また、多準位動作のために必要な狭いVt範囲を得るために、読み出し動作と類似のプログラム確認サイクルによって、プログラムの間中、閾値電圧が定期的に検

査されなければならない。本発明におけるバリスティック短チャネルサイドウォールMONOS用のプログラム確認は、プログラム電圧が極めて低く、読み出し電圧条件に非常に似ているので、従来の浮遊ゲートおよびMONOSメモリよりも単純である。

【0051】消去中の窒化物領域からの電子の除去は、窒化物領域から拡散領域へのホットホール注入によるか、あるいは窒化物領域から制御ゲートへのF－Nトンネリングによってなされる。ホットホール注入では、基板がアース接地され、拡散領域が5Vに設定され、かつ－5Vが制御ゲートに給電される。F－Nトンネリングでは、－3．5Vが基板と拡散領域の両方に給電され、5Vが制御ゲートに給電される。窒化物領域の障壁は同時に除去されなければならない。単一窒化物領域は除去されることができない。

【0052】読み出しの好ましい実施例

各窒化物領域内の2ビット多準位記憶用の読み出し動作が、0．25μプロセス用のシミュレーションに基づいて説明される。図8Aは、メモリセルおよび窒化物蓄積領域313の読み出し用の電圧条件を示す。記憶された4つの準位の閾値電圧は「11」、「10」、「01」、および「00」状態用に、それぞれ0．8V、1．2V、1．6V、および2．0Vである。このことが図8Bに示される。ワード選択ゲート用の閾値電圧は0．5Vである。読み出し中、ソース電圧は1．2Vに固定される。選択されない窒化物蓄積領域上の制御ゲートは、全ての可能な閾値状態をオーバーライドするような5Vに設定され、選択された窒化物蓄積領域上の制御ゲートは、全ての可能な閾値状態の最高閾値電圧である2．0Vに設定される。その他の制御ゲートは全てゼロに設定されて、ビット接合部はあらかじめゼロに設定される。ワード線は0Vから1．0Vに高められて、ビット接合部が監視される。

【0053】ビット接合部の感知結果は図8Cに示されるようなカーブを生じる。窒化物蓄積領域313からの読み出し中のビット線電圧感知カーブ71、73、75、および77が、それぞれ異なる閾値0．8V、1．2V、1．6V、および2．0Vと対応して示される。電圧カーブから、それぞれの状態間の電圧差が、感知マージンに良好な約300mVであることが解る。シミュレーションはまた、選択されないセルの状態が、図8Cのビット接合部電圧カーブで変化を示さないことも裏付ける。

【0054】本発明は、超短チャネルを伴ない、下層にONO窒化膜蓄積記憶領域を有するダブルサイドウォール制御ゲートを形成するための方法を提供する。拡張モードチャネルは35nm位であり、サイドウォールスペーサによって限定される。自己整合による二酸化シリコンの充填技術によって、ワードゲート間の絶縁が形成される。化学機械的研磨を用いる自己整合技術によって、ポリシリコン制御ゲートが形成される。本発明のプロセスは、2つの実施例、すなわち、バリスティック注入を有するプレーナー短チャネル構造、およびバリスティック注入を有するステップスプリットチャネル構造を含む。第3の実施例は、制御ゲート形成後の隣接ワードゲートの絶縁を提供する。

【0055】本発明を好ましい実施例によって説明してきたが、本発明の精神や範囲を逸脱することなく詳細や形式上の変更が可能なことを当業者は理解するであろう。

【0056】

【発明の効果】本発明によって、高速低電圧バリスティックプログラム、超短チャネル、超高集積度、デュアルビット多準位のフラッシュメモリが、2つまたは3つのポリシリコン・スプリット・ゲート・サイドウォール・プロセスで実現される。

【図面の簡単な説明】

【図1】　従来技術のSONOS（シリコン－酸化膜－窒化膜－酸化膜－シリコン）のデバイス構造である。

【図2A】　100nmのチャネル長のためにはソースサイド注入が高電圧動作を必要とすることを証明するような、スプリットゲート浮遊ゲートトランジスタの実験結果を示す図である。

【図2B】　40nmのチャネル長のためには、バリスティック注入動作が、はるかに低い電圧および／あるいは、はるかに高速なプログラム速度で動作することを示すような、スプリットゲート浮遊ゲートトランジスタの実験結果を示す図である。

【図3A】　超短バリスティックチャネルを有する従来のダブルサイドウォールデュアルビットスプリット浮遊ゲートセルの配列概要図である。

【図3B】　超短バリスティックチャネルを有する従来のダブルサイドウォールデュアルビットスプリット浮遊ゲートセルの配置断面図である。

【図4A】　本発明のプロセスの、第1の好ましい実施例の断面図である。

【図4B】　本発明のプロセスの、第1の好ましい実施例の断面図である。

【図4C】　本発明のプロセスの、第1の好ましい実施例の断面図である。

【図4D】　本発明のプロセスの、第1の好ましい実施例の断面図である。

【図4E】　本発明のプロセスの、第1の好ましい実施例の断面図である。

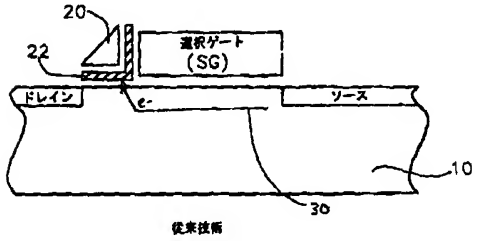
【図4F】　本発明のプロセスの、第1の好ましい実施例の断面図である。

【図4G】　本発明の完成されたメモリセルの平面図である。

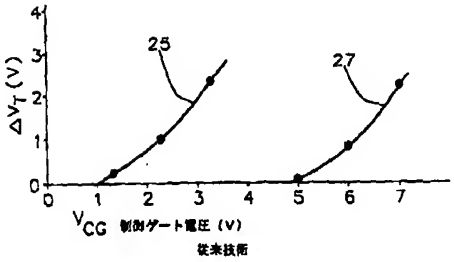
【図5B】　本発明のプロセスの、第2の好ましい実施例の断面図である。

【図5C】 本発明のプロセスの、第2の好ましい実施例の断面図である。
【図5F】 本発明のプロセスの、第2の好ましい実施例の断面図である。
【図6A】 本発明のプロセスの、第3の好ましい実施例の断面図である。
【図6B】 本発明のプロセスの、第3の好ましい実施例の断面図である。
【図6C】 本発明のプロセスの、第3の好ましい実施例の断面図である。
【図6D】 本発明のプロセスの、第3の好ましい実施例の断面図である。
【図6F】 本発明のプロセスの、第3の好ましい実施例の断面図である。
【図7A】 本発明の配置概略図である。
【図7B】 本発明の断面図である。
【図7C】 本発明において読み出し中に必要とされる電圧状態を示す。
【図8A】 本発明における読み出し中の感知電圧曲線を示す図である。
【図8B】 本発明における読み出し中の感知電圧曲線を示す図である。
【図8C】 本発明における読み出し中の感知電圧曲線を示す図である。
【符号の説明】
202……ホウ素、203……Nドーパント、204……n^+注入領域、221、234……酸化シリコン膜、235……酸化膜および／あるいは窒化膜、240……サイドウォール制御ゲート、241……シリサイド層、245……ポリシリコン、247……間隙充填材、248……ポリシリコン膜

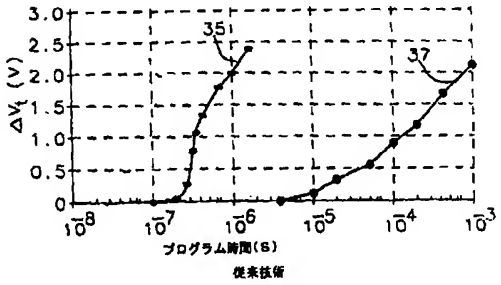
【図1】



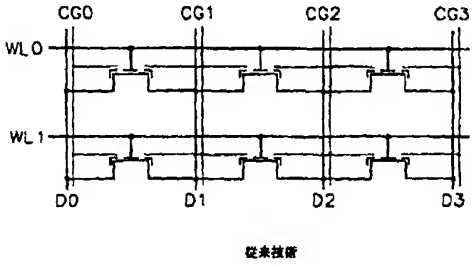
【図2A】



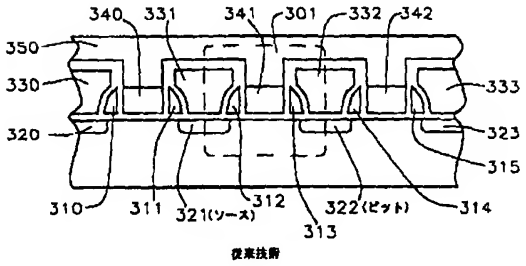
【図2B】



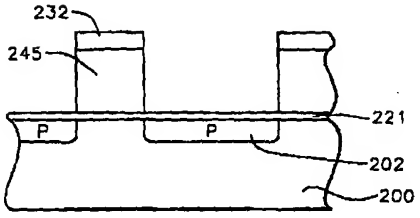
【図3A】



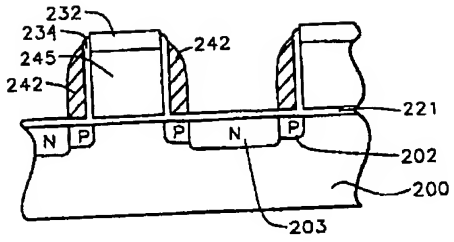
【図3B】



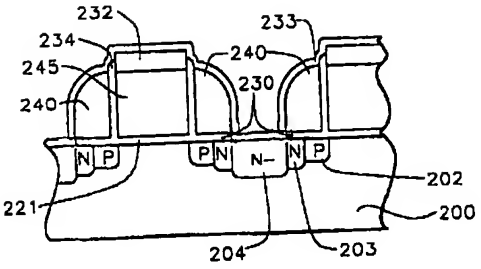
【図4A】



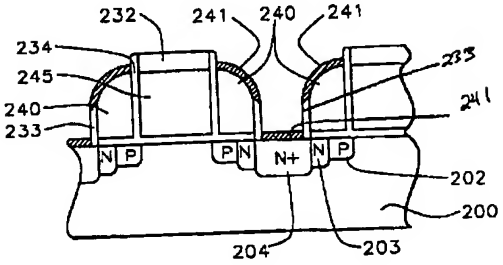
【図4B】



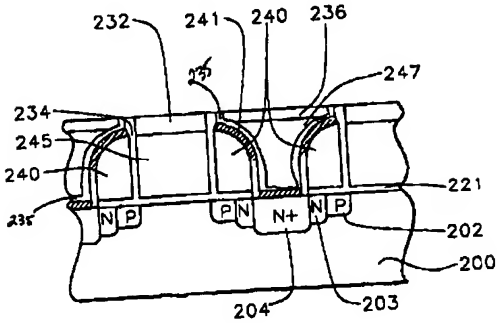
【図4C】



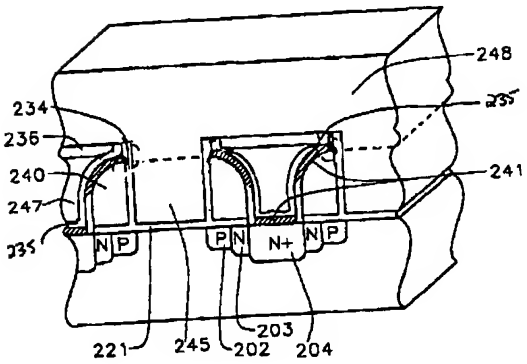
【図4D】



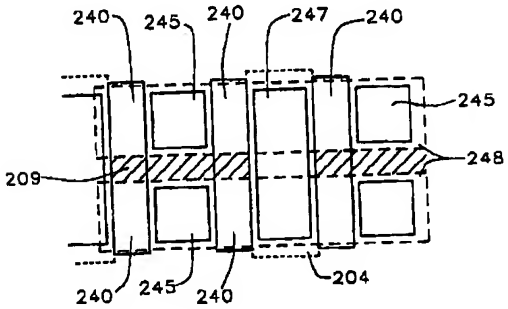
【図4E】



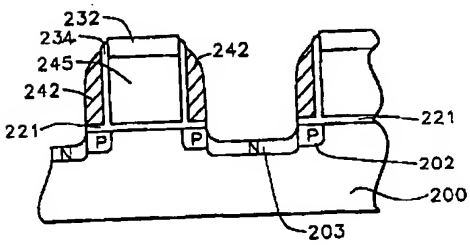
【図4F】



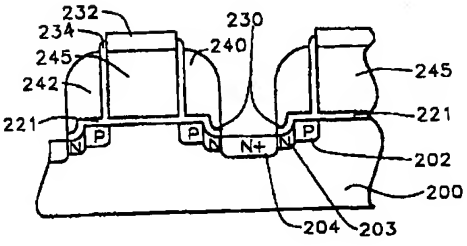
【図4G】



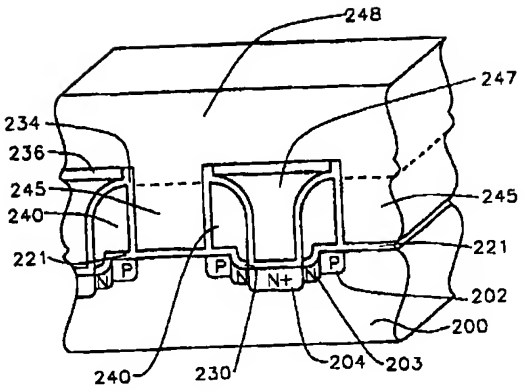
【図5B】



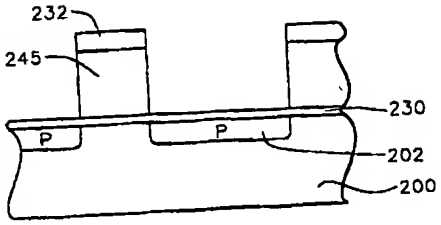
【図5C】



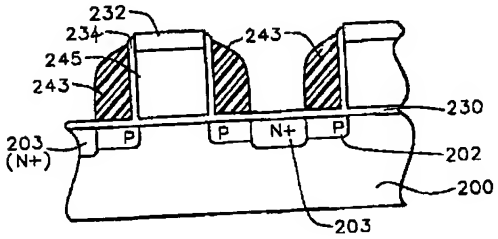
【図5F】



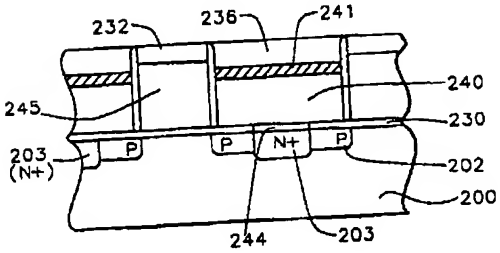
【図6A】



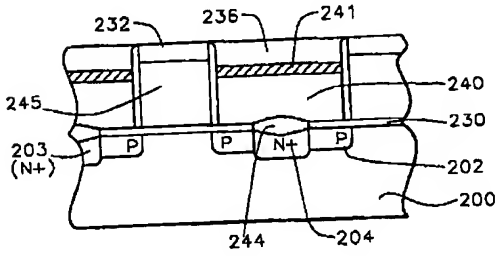
【図6B】



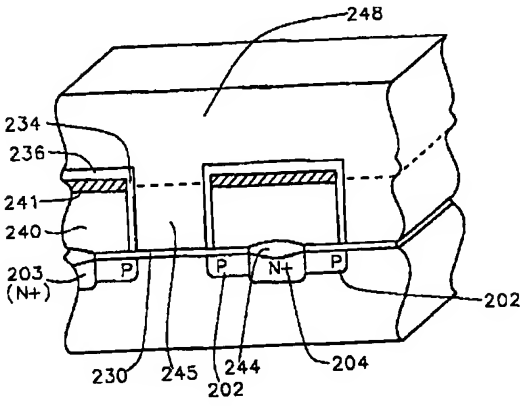
【図6C】



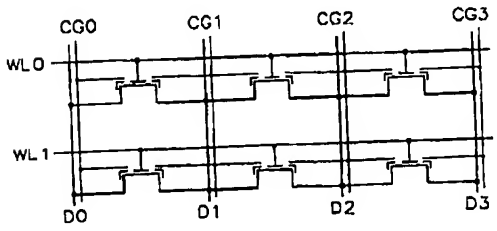
【図6D】



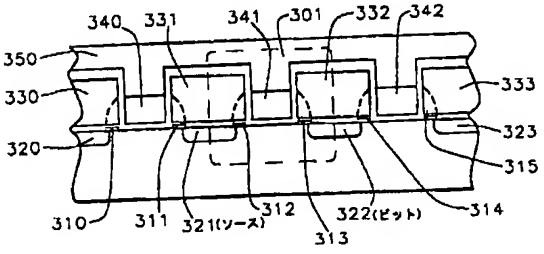
【図6F】



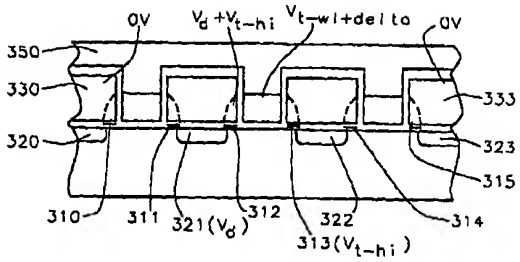
【図7A】



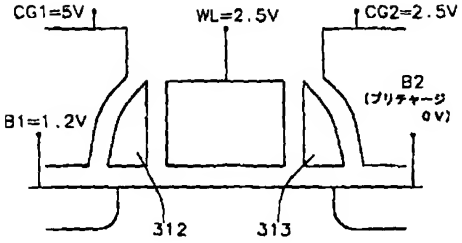
【図7B】



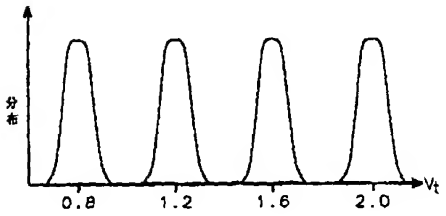
【図7C】



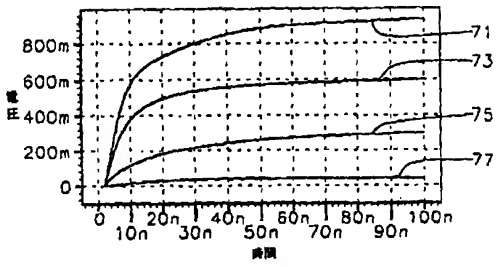
【図8A】



【図8B】



【図8C】



フロントページの続き

(72)発明者　小椋　正気
アメリカ合衆国　12590　ニューヨーク州、ワッピンガー　フォールズ、オールドホープウェルロード　140
(72)発明者　トモコ　オグラ
アメリカ合衆国　12590　ニューヨーク州、ワッピンガー　フォールズ、オールドホープウェルロード　140
(72)発明者　林　豊
茨城県つくば市梅園2丁目3番10号

Fターム(参考)　5F001 AA13 AB02 AC01 AC06 AD15
AD16 AD22 AE02 AE03 AE08
AF20 AG07 AG10 AG12
5F083 EP18 EP22 EP62 EP67 ER02
ER11 ER18 ER30 JA04 JA35
JA39 KA01 MA04 MA19 PR03
PR09 PR29 PR36 PR40 ZA21
5F101 BA45 BB02 BC01 BC11 BD05
BD06 BD14 BE02 BE05 BE07
BF05 BH09 BH14 BH19

【外国語明細書】

1. Title of Invention

PROCESS FOR MAKING AND PROGRAMMING AND OPERATING A DUAL-BIT MULTI-LEVEL BALLISTIC MONOS MEMORY

2. Claims

1. A method for fabricating a MONOS memory device comprising:

forming a gate silicon oxide layer on the surface of a semiconductor substrate;

depositing a first polysilicon layer overlying said gate silicon oxide layer;

depositing a first nitride layer overlying said first polysilicon layer;

patterning said first polysilicon layer and said first nitride layer to form word gates wherein a gap is left between two of said word gates;

forming a first insulating layer on the sidewalls of said word gates;

depositing a spacer layer overlying said word gates and said gate silicon oxide layer;

anisotropically etching away said spacer layer to leave disposable spacers on the sidewalls of said word gates;

implanting ions into said semiconductor substrate to form a lightly doped region wherein said disposable spacers act as an implantation mask;

thereafter removing said disposable spacers;

depositing a nitride-containing layer over said semiconductor substrate in said gap;

depositing a second polysilicon layer overlying said word gates and said nitride-containing layer;

anisotropically etching away said second polysilicon layer and said nitride-containing layer to leave polysilicon spacers on the sidewalls of said word gates wherein said polysilicon spacers form control sidewall spacer gates and wherein said nitride-containing layer underlying each of said control sidewall spacer gates forms a nitride region in which charge is stored ;

forming a second insulating layer on said control sidewall spacer gates;

implanting ions into said semiconductor substrate to form a bit diffusion region wherein said control sidewall spacer gates act as an implantation mask;

coating a gap filling material over the surface of said substrate wherein said gap-filling material fills said gap betwe n said two of said word gates;

planarizing said gap-filling material;

thereafter removing said first nitride layer overlying said word gates; and

depositing a third polysilicon layer overlying said substrate wherein said third polysilicon layer forms a word line connecting underlying said word gates to complete said fabrication of said MONOS memory device.

2. The method according to Claim 1 wherein said gate silicon oxide layer has a thickness of between about 5 and 10 nanometers.

3. The method according to Claim 1 wherein said first polysilicon layer is deposited by chemical vapor deposition to a thickness of between about 150 and 250 nanometers.

4. The method according to Claim 1 wherein said first nitride layer is deposited by chemical vapor deposition to a thickness of between about 50 and 100 nanometers.

5. The method according to Claim 1 wherein said first insulating layer is formed by thermally growing a silicon oxide layer to a thickness of between about 5 and 10 nanometers on the sidewalls of said word gates.

6 The method according to Claim 1 wherein said first insulating layer is formed by depositing a silicon oxide layer by chemical vapor deposition to a thickness of between about 5 and 10 nanometers on the sidewalls of said word gates.

7. The method according to Claim 1 wher in said first insulating layer is formed by d positing a silicon nitride layer to a thickn ss of betwe n about 5 and 10 nanom ters on the sidewalls of sald word gates.

8. The method according to Claim 1 wherein said first insulating layer is formed by depositing a silicon oxide layer and a silicon nitride layer to a combined thickness of between about 5 and 10 nanometers on the sidewalls of said word gates.

9. The method according to Claim 1 wherein said spacer layer comprises one of the group contalning polysilicon, plasma nitride, plasma oxynitride, and borophosphosilicate glass (BPSG) and has a thickness of between about 30 and 50 nanometers.

10. The method according to Claim 1 wherein said step of removing said disposable spacers comprises a dry chemical anisotropic etch.

11. The method according to Claim 1 wherein said step of depositing said nitride-containing layer comprises:

growing a first silicon oxide layer to a thickness of between about 3.6 and 5.0 nanometers on said semiconductor substrate;

depositing a silicon nitride layer having a thickness of about 2 to 5 nanometers overlying said first silicon oxide layer; and

depositing a second silicon oxide layer having a thickness of between about 4 and 8 nanometers overlying said silicon nitride layer.

12. The method according to Claim 1 further comprising nitriding said first silicon oxide layer before said step of depositing said sllicon nitride layer.

13. Th method according to Claim 1 wherein said sec nd p lysilicon lay r has a thickness of between about 30 and 50 nanometers.

14. The method according to Claim 1 wherein said second polysilicon layer has a thickness of between about 30 and 50 nanometers and further comprising depositing a tungsten silicide layer having a thickness of between about 60 and 100 nanometers and wherein said second polysilicon layer and tungsten silicide layer together form said control sidewall spacer gates.

15. The method according to Claim 1 wherein said second insulating layer comprises silicon oxide deposited by chemical vapor deposition to a thickness of about 10 nanometers.

16. The method according to Claim 1 wherein said second insulating layer comprises silicon nitride deposited by chemical vapor deposition to a thickness of about 10 nanometers.

17. The method according to Claim 1 further comprising:

anisotropically etching said second insulating layer to form sidewall oxide spacers on lower portions of said control sidewall spacer gates; and

thereafter siliciding upper portions of said control sidewall spacer gates and said bit diffusion region.

18. The method according to Claim 1 wherein said gap-filling material comprises one of the group containing silicon oxide and borosilicate glass.

19. The method according to Claim 1 wherein said gap-filling material comprises a conductive material and further comprising:

recessing said conductive material below the surface of said first nitride layer;

depositing a silicon oxide layer overlying said recessed conductive material; and

planarizing said silicon oxide layer wherein said conductive material and underlying said control sidewall spacer gates together form a control gate.

20. The method according to Claim 1 wherein said third polysilicon layer has a thickness of between about 150 and 200 nanometers.

21. The method according to Claim 1 further comprising siliciding said word line.

22. A method for fabricating a step split structure MONOS memory device comprising:

forming a gate silicon oxide layer on the surface of a semiconductor substrate;

depositing a first polysilicon layer overlying said gate silicon oxide layer;

depositing a first nitride layer overlying said first polysilicon layer;

patterning said first polysilicon layer and said first nitride layer to form word gates wherein a gap is left between two of said word gates;

forming a first insulating layer on the sidewalls of said word gates;

depositing a spacer layer overlying said word gates and said gate silicon oxide layer;

anisotropically etching away said spacer layer to leave disposable spacers on the sidewalls of said word gates;

etching away said gate silicon oxide layer not covered by said word gates and said disposable spacers to expose a portion of said semiconductor substrate;

etching away said exposed portion of said semiconductor substrate to form a step into said substrate;

implanting ions into said semiconductor substrate to form a lightly doped region wherein said disposable spacers act as an implantation mask;

thereafter removing said disposable spacers;

removing said gate silicon oxide layer underlying said disposable polysilicon spacers;

forming a composite layer of oxide-nitride-oxide overlying said semiconductor substrate;

depositing a second polysilicon layer overlying said word gates and said second gate silicon oxide layer;

anisotropically etching away said second polysilicon layer and said composite oxide-nitride-oxide layer to leave polysilicon spacers on the sidewalls of said word gates wherein said polysilicon spacers form sidewall control gates and wherein the nitride portion of said composite oxide-nitride-oxide layer underlying each of said sidewall control gates forms a nitride region in which charge is stored;

forming a second insulating layer on said control sidewall gates;

implanting ions into said semiconductor substrate to form a bit diffusion region wherein said control sidewall gates act as an implantation mask;

coating a gap filling material over the surface of said substrate wherein said gap-filling material fills said gap between said two of said word gates;

planarizing said gap-filling material;

thereafter removing said first nitride layer overlying said word gates; and

depositing a third polysilicon layer overlying said substrate wherein said third polysilicon layer forms a word line connecting underlying said word gates to compl te said fabrication of said MONOS memory device.

23. The method according to Claim 22 wherein said first polysilicon layer is deposited by chemical vapor deposition to a thickness of between about 150 and 250 nanometers.

24. The method according to Claim 22 wherein said first nitride layer is deposited by chemical vapor deposition to a thickness of between about 50 and 100 nanometers.

25. The method according to Claim 22 wherein said first insulating layer is formed by thermally growing a silicon oxide layer to a thickness of between about 5 and 10 nanometers on the sidewalls of said word gates.

26. The method according to Claim 22 wherein said first insulating layer has a thickness of between about 5 and 10 nanometers on the sidewalls of said word gates.

27. The method according to Claim 22 wherein said spacer layer comprises one of the group containing polysilicon, plasma nitride, plasma oxynitride, and borophosphosilicate glass (BPSG) and has a thickness of between about 30 and 50 nanometers.

28. The method according to Claim 22 wherein said step of removing said disposable spacers comprises a dry chemical anisotropic etch.

29. The method according to Claim 22 wherein said step into said semiconductor substrate has a depth of between about 20 and 50 nanometers.

30. The method according to Claim 22 after said tep of removing said gate silicon oxide layer underlying said disposable spacers further comprising rounding the corners of said step.

31. The method according to Claim 30 wherein said step of rounding said corners of said step comprises a rapid thermal anneal at between about 1000 and 1100 °C for about 60 seconds.

32. The method according to Claim 30 wherein said step of rounding said corners of said step comprises annealing in hydrogen at about 900 °C at a pressure of between about 200 and 300 mtorr.

33. The method according to Claim 22 wherein said oxide-nitride-oxide composite layer comprises:

a first silicon oxide layer having a thickness of between about 3.6 and 5.0 nanometers;

a second silicon nitride layer having a thickness of about 2 to 5 nanometers; and

a third silicon oxide layer having a thickness of between about 4 and 8 nanometers.

34. The method according to Claim 22 wherein said second polysilicon layer has a thickness of between about 30 and 50 nanometers.

35. The method according to Claim 22 wherein said second polysilicon layer has a thickness of between about 30 and 50 nanometers and further comprising depositing a tungsten silicide layer having a thickness of between about 60 and 100 nanometers and

wherein said third polysilicon layer and tungsten silicide layer together form said control sidewall spacer gates.

36. The method according to Claim 22 wherein said second insulating layer comprises silicon oxide deposited by chemical vapor deposition to a thickness of about 10 nanometers.

37. The method according to Claim 22 wherein said second insulating layer comprises silicon nitride deposited by chemical vapor deposition to a thickness of about 10 nanometers.

38. The method according to Claim 22 further comprising:

anisotropically etching said second insulating layer to form sidewall oxide spacers on lower portions of said control sidewall spacer gates; and

thereafter siliciding upper portions of said control sidewall spacer gates and said bit diffusion region.

39. The method according to Claim 22 wherein said gap-filling material comprises one of the group containing silicon oxide and borosilicate glass.

40. The method according to Claim 22 wherein said gap-filling material comprises a conductive material and further comprising:

recessing said conductive material below the surface of said first nitride layer;

depositing a silicon oxide layer overlying said recessed conductive material; and

planarizing said silicon oxide layer wherein said conductive material and underlying said control sidewall spacer gates together form a control gate.

41. The method according to Claim 22 wherein said third polysilicon layer has a thickn ss of between about 90 and 180 nanometers.

42. The method according to Claim 22 further comprising siliciding said word line.

43. The method according to Claim 22 further comprising siliciding said word line.

44. A method for fabricating a MONOS memory device comprising:

forming a nitride-containing layer on the surface of a semiconductor substrate;

depositing a first polysilicon layer overlying said nitride-containing layer;

depositing a second nitride layer overlying said first polysilicon layer;

patterning said first polysilicon layer and said second nitride layer to form word gates wherein a gap is left between two of said word gates;

forming a first insulating layer on the sidewalls of said word gates;

depositing a spacer layer overlying said word gates and said gate silicon oxide layer;

anisotropically etching away said spacer layer to leave disposable spacers on the sidewalls of said word gates;

implanting ions into said semiconductor substrate to form a bit diffusion junction wherein said disposable spacers act as an implantation mask;

thereafter removing said disposable spacers;

depositing a second polysilicon layer overlying said word gates and filling said gap;

recessing said second polysilicon layer below a surface of said second nitride layer;

silicidlng said recessed second polysillcon layer wherein said silicided recessed second polysilicon layer forms a control gate;

depositing an oxide layer overlying said silicided recessed second polysilicon layer;

thereafter removing said second nitride layer overlying said word gates; and

depositing a third polysilicon layer overlying said substrate wherein said third polysilicon layer forms a word line connecting underlying said word gates to complete said fabrication of said MONOS memory device.

45. The method according to Claim 44 wherein said step of forming said nitride-containing layer comprises:

growing a first silicon oxide layer to a thickness of between about 3.6 and 5.0 nanometers on said semiconductor substrate;

depositing a silicon nitride layer having a thickness of about 2 to 5 nanometers overlying said first silicon oxide layer; and

depositing a second silicon oxide layer having a thickness of between about 4 and 8 nanometers overlying said silicon nitride layer.

46. The method according to Claim 45 further comprising nitriding said first silicon oxide layer before said step of depositing said silicon nitride layer.

47. The method according to Claim 44 wherein said first polysilicon layer is deposited by chemical vapor deposition to a thickness of between about 150 and 250 nanometers.

48. The method according to Claim 44 wherein said second nitride layer is deposited by chemical vapor deposition to a thickness of between about 50 and 100 nanometers.

49. The method according to Claim 44 wherein said first insulating layer is formed to a thickness of between about 5 and 10 nanometers on the sidewalls of said word gates.

50. The method according to Claim 44 wherein said spacer layer comprises one of the group containing polysilicon, plasma nitride, plasma oxynitride, and borophosphosilicate glass (BPSG) and has a thickness of between about 30 and 50 nanometers.

51. The method according to Claim 44 further comprising before said step of removing said disposable spacers:

etching away said second silicon oxide layer not covered by said disposable spacers;

depositing a third silicon oxide layer overlying said nitride layer to a thickness of between about 4 and 6 nanometers; and

oxidizing said third silicon oxide layer to form an oxide layer having a thickness of about 20 nanometers over said nitride layer whereby coupling capacitance between said control gate and said bit diffusion is reduced.

52. The method according to Claim 44 wherein said step of removing said disposable spacers comprises a dry chemical anisotropic etch.

53. The method according to Claim 44 wherein said second polysilicon layer has a thickness of between about 30 and 50 nanometers.

54. The method according to Claim 44 wherein said second insulating layer comprises silicon oxide deposited by chemical vapor deposition to a thickness of about 10 nanometers.

55. The method according to Claim 44 wherein said second insulating layer comprises silicon nitride deposited by chemical vapor deposition to a thickness of about 10 nanometers.

56. The method according to Claim 44 wherein said third polysilicon layer has a thickness of between about 150 and 200 nanometers.

57. A method for fabricating a flash memory device comprising:

providing word gates overlying a gate silicon oxide layer on the surface of a semiconductor substrate wherein a gap is left between two of said word gates;

forming disposable spacers on the sidewalls of said word gates;

implanting ions into said semiconductor substrate to form a lightly doped region wherein said disposable spacers act as an implantation mask;

thereafter removing said disposable spacers;

forming sidewall polysilicon gates on the sidewalls of said word gates, each of said sidewall polysilicon gates having an underlying nitride-containing layer wherein the nitride region of said nitride-containing layer acts as a nitride charge region;

implanting ions into said semiconductor substrate to form a bit diffusion region wherein said sidewall polysilicon gates act as an implantation mask;

forming an insulating layer on said sidewall gates;

filling said gap between said two of said word gates with a second polysilicon layer;

recessing said second polysilicon layer;

siliciding said recessed second polysilicon layer;

covering said silicided recessed second polysilicon layer with an oxide layer wherein said silicided recessed second polysilicon layer along with underlying said sidewall polysilicon gates form a control gate; and

depositing a third polysilicon layer overlying said substrate wherein said third polysilicon layer forms a word lin connecting said word gates to complete said fabrication of said flash memory device.

58. The method according to Claim 57 wherein said first polysillcon layer has a thickness of between about 150 and 250 nanometers.

59. The method according to Claim 57 wherein said disposable spacers comprise one of the group containing polysilicon, plasma nitride, plasma oxynitride, and borophosphosilicate glass (BPSG).

60. The method according to Claim 57 wherein said nitride-containing layer comprises a first layer of silicon oxide, a second layer of sillcon nitride, and a third layer of silicon oxide.

61. The method according to Claim 57 after said step of removing said disposable spacers further comprising etching into said semiconductor substrate to form a step into said semiconductor substrate having a depth of between about 20 and 50 nanometers.

62. The method according to Claim 57 further comprising rounding the corners of said step.

63. The method according to Claim 62 wherein said step of rounding said corners of said step comprises a rapid thermal anneal at between about 1000 and 1100 °C for about 60 seconds.

64. The method according to Claim 62 wherein said step of rounding said corners of said step comprises annealing in hydrogen at about 900 °C at a pressure of between about 200 and 300 mtorr.

65. The method according to Claim 57 wherein a channel length defined from an edge of said word gate to an edge of adjacent said bit diffusion region is between about 30 and 50 nm and whereby ballistic electron injection occurs.

66. A MONOS memory cell comprising:

a word gate on the surface of a semiconductor substrate;

sidewall control gates on sidewalls of said word gate, separated from said word gates by an insulating layer;

nitride regions within an ONO layer underlying said sidewall control gates wherein electron memory storage is performed within said nitride regions;

a polysilicon word line overlying and connecting said word gate with word gates in other said memory cells and overlying said sidewall control gates, separated from said sidewall control gates by an insulating layer; and

bit line diffusions within said semiconductor substrate adjacent to each of said sidewall control gates.

67. The MONOS memory cell of Claim 66 wherein each sidewall control gate is separated from a sidewall control gate of another said memory cell by an insulating layer.

68. The MONOS memory cell of Claim 66 wherein each control gate comprises a polysilicon layer between two of said word gates overlying said bit diffusion region and said sidewall control gates wherein said nitride regions underlie only said sidewall control gates.

69. The MONOS memory call of Claim 66 wherein a channel length defined from an edge of said w rd gate to an edge of adjacent said bit diffusion region is between about 30 and 50 nm and whereby ballistic electron injection occurs.

70. The MONOS memory cell of Claim 66, wherein one of said nitride regions is a selected nitride region, and the other of said nitride regions is an unselected nitride region, and wherein said bit line diffusion near said selected nitride region is a bit diffusion, and said bit line diffusion near said unselected nitride region is a source diffusion, wherein a read operation of said cell is performed by:

over-riding said unselected nitride region;

providing a voltage on said word gate having a sum of the word gate threshold voltage, an overdrive voltage, and the voltage on said source diffusion;

providing a voltage on said control gate adjacent to said selected nitride region sufficient to allow for reading of the selected nitride region; and

reading said cell by measuring the voltage level on said bit diffusion.

71. The MONOS memory cell of Claim 70 wherein said memory cell is one of many cells in a MONOS memory array, and further comprising applying a control gate voltage of 0 volts to all cells beside the cell desired to be read.

72. The MONOS memory cell of Claim 70 wherein said memory cell is one of many cells in a MONOS memory array, and further comprising applying a control gate voltage of –0.7 volts to all cells beside the cell desired to be read in order to stop leakage.

73. The MONOS memory cell of Claim 66 wherein the voltage level on said bit diffusion may represent one of multiple threshold levels of said cell.

74. The MONOS memory cell of Claim 66, wherein one of said nitride regions is a selected nitride region, and the other of said nitride regions is an unselected nitride region, and wherein said bit line diffusion near said selected nitride region is a bit diffusion, and said bit line diffusion near said unselected nitride region is a source diffusion, wherein a program operation of said cell is performed by:

providing a high voltage on said unselected control gate to over-ride said unselected nitride region;

raising the control gate voltage of said selected nitride region;

providing a fixed voltage on said bit diffusion;

providing a voltage on said word line which is greater than said word gate threshold voltage; and

lowering the voltage of said source diffusion such that current flows from said source diffusion to said bit diffusion wherein ballistic injection of electrons occurs from a channel region to said selected nitride region when current flows.

75. The MONOS memory cell of Claim 74 wherein multiple thresholds can be programmed by varying said voltage on said bit diffusion line.

76. The MONOS memory cell of Claim 74 wherein said memory cell is one of many cells in a MONOS memory array, and further comprising disabling nitride regions in adjacent cells sharing a word line by applying a control gate voltage of 0 volts to said adjacent cells.

77. The MONOS memory cell of Claim 66, wherein one of said control gates is a selected control gat and its underlying nitride region is a selected nitride region, and the other of said control gates is an unselected control gate and its underlying nitride region is an unselected nitride region, and wherein said bit line diffusion near said selected nitride region is a bit diffusion, and said bit line diffusion near said unselected nitride region is a source diffusion, wherein a program operation of said cell is performed by:

providing a high voltage on said unselected control gate to over-ride said unselected nitride region; and

varying a voltage on said selected control gate.

78. The MONOS memory cell of Claim 66 wherein said memory cell is one of many cells in a flash memory array that share a word line, and further comprising simultaneously programming several of said cells with different threshold levels by varying the voltage either of said control gate or said bit diffusion.

79. The MONOS memory cell of Claim 66, wherein an erase operation of a block of nitride regions is performed by:

providing a positive voltage to said bit line diffusions; and

providing a negative voltage to said control gates over said bit line diffusions.

80. The MONOS memory cell of Claim 66, wherein an erase operation of a block of nitride regions is performed by:

providing a negative voltage to said semiconductor substrate and to said bit line diffusions; and providing a positive voltage to said control gates.

81. A method of reading a MONOS memory cell, wherein the MONOS memory cell comprises:

a word gate on the surface of a semiconductor substrate;

sidewall control gates on sidewalls of said word gate, separated from said word gates by an insulating layer;

nitride regions within an ONO layer underlying said sidewall control gates wherein electron memory storage is performed within said nitride regions;

a polysilicon word line overlying and connecting said word gate with word gates in other said memory cells and overlying said sidewall control gates, separated from said sidewall control gates by an insulating layer; and

bit line diffusions within said semiconductor substrate adjacent to each of said sidewall control gates.

wherein one of said nitride regions is a selected nitride region, and the other of said nitride regions is an unselected nitride region, and wherein said bit line diffusion near said selected nitride region is a bit diffusion, and said bit line diffusion near said unselected nitride region is a source diffusion,

wherein a read operation of said cell is performed by:

over-riding said unselected nitride region;

providing a voltage on said word gate having a sum of the word gate threshold voltage, an overdrive voltage, and the voltage on said source diffusion;

providing a voltage on said control gate adjacent to said selected nitride region sufficient to allow for reading of the selected nitride region; and

reading said cell by measuring the voltage level on said bit diffusion.

82. A method of programming a MONOS memory cell, wherein said MONOS memory cell comprises:

a word gate on the surface of a semiconductor substrate;

sidewall control gates on sidewalls of said w rd gate, separated from said word gates by an insulating layer;

nitride regions within an ONO layer underlying said sidewall control gates wherein electron memory storage is performed within said nitride regions;

a polysilicon word line overlying and connecting said word gate with word gates in other said memory cells and overlying said sidewall control gates, separated from said sidewall control gates by an insulating layer; and

bit line diffusions within said semiconductor substrate adjacent to each of said sidewall control gates;

wherein one of said control gates is a selected control gate and its underlying nitride region is a selected nitride region, and the other of said control gates is an unselected control gate and its underlying nitride region is an unselected nitride region, and wherein said bit line diffusion near said selected nitride region is a bit diffusion, and said bit line diffusion near said unselected nitride region is a source diffusion,

wherein said method of programming the cell comprises the steps of:

providing a high voltage on said unselected control gate to over-ride said unselected nitride region; and

varying a voltage on said selected control gate.

83. A method of erasing a MONOS memory cell, wherein said MONOS memory cell comprises:

a word gate on the surface of a semiconductor substrate;

sidewall control gates on sidewalls of said word gate, separated from said word gates by an insulating layer;

nitride regions within an ONO layer underlying said sidewall control gates wherein electron m mory storag is performed within said nitride regions;

a polysilicon word line overlying and connecting said word gate with word gates in other said memory cells and overlying said sidewall control gates, separated from said sidewall control gates by an insulating layer; and

bit line diffusions within said semiconductor substrate adjacent to each of said sidewall control gates;

wherein said method of erasing a block of said nitride regions comprises the steps of:

providing a positive voltage to said bit line diffusions; and

providing a negative voltage to said control gate over said bit line diffusions.

84. A flash memory device comprising:

word gates on the surface of a semiconductor substrate;

sidewall control gates on the sidewalls of said word gates separated from said word gates by an insulating layer;

bit line diffusions within said semiconductor substrate between two of said sidewall control gates; and

nitride charge regions underlying said sidewall control gates.

85. The device according to Claim 84 further comprising:

an insulating layer overlying said sidewall control gates; and

a word line overlying said control gates and connecting said word gates.

86. The MONOS memory call of Claim 84 wherein a channel length defined from an edge of said word gate to an edge of adjacent said bit diffusion region is between about 30 and 50 nm and whereby ballistic electron injection occurs.

3. Detailed Explanation of the Invention

Field of Invention

The invention relates to methods of forming high-density Metal/polysilicon Oxide Nitride Oxide Silicon (MONOS) memory arrays and the resulting high density MONOS memory arrays.

Description of Prior Art

Floating gate and MONOS are two types of non-volatile memories. In conventional floating gate structures, electrons are stored onto a floating gate, by either F-N tunneling or source side injection. Conventional MONOS devices store electrons usually by direct tunneling in the Oxide-Nitride-Oxide (ONO) layer which is below the memory word gate. Electrons are trapped in the Nitride layer of the ONO composite. The MONOS transistor requires one less polysilicon layer than the floating gate device, which simplifies the process and could result in a denser array.

MONOS structures are conventionally planar devices in which an ONO composite layer is deposited beneath the word gate. The thickness of the bottom oxide of the ONO layer is required be less than 3.6nm, in order to utilize direct tunneling for program operations. However in 1998, a MONOS structure with a bottom oxide thickness of 5.0nm, and side wall polysilicon gates and source side injection program was first reported by Kuo-Tung Chang et al, in, "A New SONOS Memory Using Source Side Injection for Programming", IEEE Electron Letters, Vol.19, No. 7, July 1998. In this structure, as shown in Fig. 1, a side wall spacer

20 is formed on one side of th word gate by a typical side wall proc ss, and the ONO c mposite 22 is underneath the side wall gate, instead of under the word gate as for conventional MONOS memory cells. The channel under the SONOS side wall control gate is larger than 100nm, so the program mechanism is source side injection, which is faster and requires lower voltages than electron tunneling, despite the thicker bottom oxide. During source side injection, a channel potential is formed at the gap between the side wall gate and the select/word gate. Channel electrons 30 are accelerated in this gap region and become hot enough to inject into the ONO layer. Thus Kuo-Tung Chang's SONOS memory is able to achieve better program performance than previous direct tunneling MONOS cells.

While the SONOS memory cell is unique among MONOS memories for its split gate structure and source side injection program, its structure and principles of program are similar to those for a conventional split gate floating gate device. Both cell types have a word gate and side wall spacer gate in series. The most significant differences lie in the manner of side wall gates utilization and electron storage regions. In the split gate floating gate cell, the side wall spacer is a floating gate onto which electrons are stored. The floating gate voltage is determined by capacitance coupling between the word gate, diffusion, and floating gate. For the SONOS cell, electrons are stored in the nitride region beneath the side wall spacer, which is called the control gate. The nitride region voltage is directly controlled by the voltage of the above side wall gate.

A floating gate memory cell having faster program and higher density was introduced in co-pending U.S. Patent Application Serial Number 09/313,302 to the same inventors, filed on May 17, 1999. Fig. 3A is an array schematic and Fig. 3B is a layout cross-section of this fast program, dual-bit, and high density memory cell. In this memory structure, high density is achieved by pairing two side wall floating gates to one word gate (for example, floating gates 312 and 313 and word gate 341), and sharing interchangeable source-drain diffusions (321 and 322)

between cells. Thus a single memory cell has two sites of electron storage. Additional polysilicon lines "control gates" run in parallel to the diffusions and orthogonal to the word gates. The control gates (331 and 332) couple to the floating gates and provide another dimension of control in order to individually select a floating gate from its pair. This memory is further characterized by fast programming due to ballistic injection. Using the same device structure, if the side wall gate channel is reduced to less than 40nm with proper impurity profiles, the injection mechanism changes from source side injection to a new and much more efficient injection mechanism called balllistic injection. The balllstic injection mechanism has been proven by S. Ogura in "Step Split Gate Cell with Ballistic Direction Injection for EEPROM/Flash", IEDM 1998, pp.987. In Figure 2A, results between ballistic injection (line 25) and conventional source side injection (line 27) are compared for a floating gate memory cell. Although the structures are very similar, when the control gate is 100nm, the injection mechanism is source side injection. However, as illustrated in Fig. 2B, when the channel is reduced to 40nm to satisfy the short channel length requirement for ballistic injection (line 35), program speed increases by three orders of magnitude under the same bias conditions, or at half of the floating gate voltage requirement for source side injection (line 37).

In contrast, the side wall channel length of Kuo Tung Chang's SONOS memory structure is 200nm, so the program mechanism is source side injection. Thus there is a significant dependence between the short channel length and the injection mechanism.

SUMMARY OF THE INVENTION

In this invention, a fast low voltage ballistic program, ultra-short channel, ultra-high density, dual-bit multi-level flash memory is achieved with a two or three polysilicon split gate side wall process. The structure and operation of this invention is enabled by a twin MONOS cell structure having an ultra-short control gate channel of less than 40nm, with ballistic injection

(S. Ogura) which provides high electron injection efficiency and very fast program at low program voltages of 3~5V. The cell structure is realized by (i) placing side wall control gates over a composite of Oxide-Nitride-Oxide (ONO) on both sides of the word gate, and (ii) forming the control gates and bit diffusion by self-alignment and sharing the control gates and bit diffusions between memory cells for high density. Key elements used in this process are:

(i) Disposable side wall process to fabricate the ultra short channel and the side wall control gate with or without a step structure.

(ii) Self-aligned definition of the control gate over the storage nitride and the bit line diffusion, which also runs in the same direction as the control gate.

The features of fast program, low voltage, ultra-high density, dual-bit, multi-level MONOS NVRAM of the present invention include:

1. Electron memory storage in nitride regions within an ONO layer underlying the control gates.
2. High density dual-bit cell in which there are two nitride memory storage elements per cell
3. High density dual-bit cell can store multi-levels in each of the nitride regions
4. Low current program controlled by the word gate and control gate
5. Fast, low voltage program by ballistic injection utilizing the controllable ultra-short channel MONOS
6. Side wall control poly gates to program and read multi-levels while masking out memory storage state effects of the unselected adjacent nitride regions and memory cells.

The ballistic MONOS memory cell is arranged in the following array: each memory cell contains two nitride regions for one word gate, and ½ a source diffusion and ½ a bit diffusion. Control gates can be defined separately or shared together over the same diffusion.

Diffusions are shared between cells and run in parallel to the side wall control gates, and perpendicular to the word lin .

A summary of the operating conditions for multi-level storage is given in Figure 38. During read, the following conditions need to be met: the voltage of the unselected control gate within a selected memory cell must be greater than the threshold voltage of the control + source voltage. The word select gate in the control gate pair is raised to the threshold voltage of the word gate + an override delta of around 0.5V + source voltage (Vt-wl + Voverdrive +Vs). Un-selected MONOS cells will be disabled by reducing the associated control gates to 0V. Program conditions are: Word line voltage is greater than threshold + an overdrive voltage delta for low current program. Both control gates in the selected pair are greater than Vt-high (the highest threshold voltage within the range of multi-level thresholds) + override delta. Adjacent memory cells sharing the same word line voltage are disabled by adjusting the control gates only.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Presented in this invention is a fabrication method for a ballistic twin MONOS memory cell with two nitride memory elements and two shared control gates. The method can be applied to a device with a flat channel and/or a device having a step channel under the nitride layer in the MONOS cell.

The procedures for formation of shallow trench isolation, p-well, and n-well are the same as for conventional CMOS processing and will not be shown. The polysilicon word gate is also defined by conventional CMOS processing as shown in Fig. 4A. In order to define the word gate, the memory gate silicon oxide 221 is formed to a thickness of between about 5 and 10 nanometers. Then the polysilicon 245 with a thickness of between about 150 and 250 nm for the gate material is deposited by chemical vapor deposition (CVD). A nitride layer 232 is deposited by CVD to a thickness of between about 50 and 100nm to be used later as an etch stop layer for chemical mechanical polishing (CMP). Normal CMOS processing defines the memory word gates; i.e., photoresist and masking processes with exposure, development, and vertical etching of the nitride 232 and polysilicon 245 by reactive ion etching (RIE) are performed. Extra boron 202 is ion implanted at low energy (less than about 10KeV energy) with an ion dosage of between 3E12 to 3E13 ions per cm^2, in order to adjust VT under the floating gate. After removing the photoresist which was used to define the word gate, the word gate is obtained as shown in Fig 4A.

A thin silicon oxide layer 234 of between about 5 and 10 nm can be thermally grown on the side wall polysilicon, or SiO_2 and/or SiN film can be deposited by uniform CVD, as shown in Fig 4B. Then the disposable sidewall process, which defines a controllably short channel and provides fast programming by high electron injection efficiency, is performed. A thin polysilicon layer typically having a thickness of between about 30 to 50 nm is deposited. Then a vertical or anisotropic polysilicon etch is performed, which forms the disposable sidewall spacer 242 on both sides of the word gate 245, as shown in Fig.4B. Implantation with an N dopant 203 such as arsenic is performed with an ion dosage of between 3E13 and 4E13/cm^2 at 10 to 15KeV. Thus, the thickness of the polysilicon layer determines the effective channel length under the control gate.

Referring now to Fig. 4C, the disposable side wall spacer 242 is gently removed by a dry chemical anisotropic etch. A typical etch ambient for this step is $HBr/Cl_2/O_2$. The bottom silicon oxide 221 is then gently etched out by buffered (with for example water of ammonium hydroxide) hydrofluoric acid (BHF), Vapor HF, or a reactive ion etch such as CF_2/O_2. A composite layer of oxide-nitride-oxide 230 is formed. Layer 230 is shown without the three layers for simplicity. The bottom oxide is thermally grown and the thickness is between 3.6 and 5 nm, which is slightly thicker than the limit of direct tunneling (3.6nm), the silicon nitride layer deposited by chemical vapor deposition is about 2 to 5 nm, and the top oxide is deposited by CVD deposition and is between about 4 and 8 nm. Thermal oxidation may be added to improve the top oxide quality. Also short nitridation in an N_2O environment can be added to improve the bottom oxide reliability prior to the deposition of the nitride layer.

Now, an insitu phosphorus-doped thin polysilicon layer between about 30 and 50 nm and tungsten silicide between 60 and 100 nm is deposited by CVD. The composite

layer of polysilicon and tungsten silicide becomes the control sid wall spacer gate. A vertical, anisotropic reactive etch is performed to form the sidewall control gate 240, as shown in Fig. 4C. The composite oxide-nitride-oxide layer is also etched through, leaving this ONO layer 230 only underlying the sidewall control gates.

A thin CVD of silicon oxide or nitride 233 with a thickness of about 10nm is deposited. Phosphorus and /or Arsenic for n+ junction 204 is implanted subsequently, at a dosage of between 3E14 to 5E15 ions per cm^2, as shown in Fig.4C. The total thickness is between 90 to 150 nm, which is equal to the summation of effective control gate channel length and lateral out diffusion of the n+ junction.

As an option, the sidewall spacer gate 240 can be simply an insitu phosphorus or As doped polysilicon layer instead of the composite layer of polysilicon and tungsten silicide. After the formation of the n+ junction and the deposition of a thin CVD of silicon oxide or nitride 233 with a thickness of about 10nm, the vertical reactive ion etch is performed to form sidewall oxide spacer 233 on the gate 240 when the control gate requires low resistivity and silicidation, as shown in Fig. 4D. In typical silicidation, about 10 nm Co or Ti is deposited by plasma sputtering and a Rapid Thermal Anneal at about 650 °C is performed. The formation of silicide layer 241 on the top part of gate 240 and diffusion 204 are shown in Fig 4D. Although silicidation 241 is shown in Fig. 4D, it is not required. It is an option to reduce the RC time constant of the control gate lines or diffusion lines in order to improve performance in all modes of operation, read, program, and erase.

An oxide and/or nitride layer 235 for contamination barrier is deposited by CVD. Then a layer of CVD silicon oxide or BSG 247 is deposited to fill the gap. The gap fill material is polished by CMP up to the nitride layer 232.

As an option, the gap fill material 247 can be a conductive material like polysilicon or W, which can be used for reducing the RC time constant of the sidewall gat or bit diffusion depending on the need. When the conductive layer is polished by CMP up to the nitride layer 232, the conductive layer is several hundred nanometers (50nm) recessed by vertical reactive ion etch. Then a CVD SiO_2 layer (about 50nm) is deposited and CMP is performed as illustrated by 236 as shown in Fig 4E.

The nitride layer 232 in Fig. 4E is selectively etched by H_3PO_4 or etched by a chemical dry etch. The polysilicon layer thickness of between 150 and 200 nm is deposited by CVD. This polysilicon layer 248 and the underlying polysilicon word gate 245 are defined by normal photoresist and RIE processes. The structure at this point is as shown in Fig. 4F.

The polysilicon layer 248 acts as a word line wire by connecting adjacent word line gates. The final memory cell is completed at this point. This word polysilicon layer can be silicided with Ti or Co to reduce the sheet resistance. A typical bird's-eye view of the memory cell is shown in Fig 4G. The shallow trench isolation region is shown by area 209 in Fig. 4G.

The preceding processes describe fabrication of planar channel floating gates with very short channel (30 to 50nm). By modifying and adding a few process steps, a step split structure with more efficient ballistic injection can be fabricated using the same process integration scheme as the planar structure. This second embodiment of the present invention will be described with reference to Figs. 5B, 5C, and 5F.

After forming disposable sidewall spacer 242 by etching vertically the doped polysilicon, the silicon oxide layer 221 is vertically etched which corresponds to Fig 4B. In order to

form a step split memory cell, th deviation starts at this point by continuing to etch into the silicon substrate by approximately 20 to 50nm. Then the bottom of the step is lightly implanted with Arsenic to form N-region 203 using the poly sidewall as a mask as shown in Fig 5B, where the dosage is about 3E13 to 4E13/cm^2 at 10 to 15KeV. Next, the N+ doped polysilicon disposable spacer is selectively removed by a wet etch (HNO_3/HF/Acitic acid, or H_3PO_4 or NH_4OH) or a dry plasma etch to the lightly doped bulk N- region. The bulk etching during this disposable spacer etch can be included as part of step etching. After gently etching off the left over gate oxide 221 under the disposable polysilicon spacer, the silicon surface is cleaned. The total step into silicon should be about 20 to 50 nm. If the step corner is sharp, corner rounding by rapid thermal anneal (RTA) at between about 1000 to 1100° C for about 60 seconds can be added as an option or a hydrogen anneal at 900°C and at a pressure of 200 to 300 mtorr can be performed. After these modifications and additions, the fabrication sequence returns to the procedures described previously.

Referring to Fig. 5C, a composite layer of oxide-nitride-oxide is formed. Layer 230 is shown without the three layers for simplicity. The bottom oxide is thermally grown and the thickness is between 3.6 and 5 nm, which is slightly thicker than the limit of direct tunneling (3.6nm), the silicon nitride layer deposited by chemical vapor deposition (CVD) is about 2 to 5 nm, and the top oxide is deposited by CVD deposition and is between about 4 and 8nm. Thermal oxidation may be added to improve the top oxide quality. Also, short nitridation in an N_2O environment can be added to improve the bottom oxide reliability prior to the deposition of the nitride layer.

Then an insitu phosphorous-doped polysilicon layer, which becomes the control gate, is deposited having a thickness of between 90 to 180 nm, and a vertical or anisotropic polysilicon etch is performed to form the sidewall gate 240, a shown in Fig 5C. By

following the process st ps given for the planar split device, the step-split device can be fabricat d as shown in Fig 5F. This sidewall polysilicon gate can be silicided or replaced by refractory silicide as utilized in the first embodiment of the flat channel MONOS twin cell.

In the above process steps for both the planar and step devices, the disposable side wall spacer 242 can be plasma nitride or oxynitride or Boron Phosphorus Silicate Glass (BPSG) instead of polysilicon, since the etching rate of that material to the thermal silicon oxide can be very high (for example at least 10-100 times) in H_3PO_4 acid or diluted HF.

A third embodiment of the present invention will be described with reference to Figs. 6A-6D and 6F. The third embodiment of the present invention will be a simplified process of the first embodiment of the planar twin MONOS memory cell with a slight program speed penalty because controllability will be lost due to the usage of a single large spacer instead of two side wall spacers. Deviation from the normal CMOS process starts prior to deposition of word gate polysilicon 245. A composite layer of oxide-nitride-oxide (ONO), 230 in Fig 6A, is formed. Layer 230 is again shown without the three layers for simplicity. The bottom silicon oxide layer is preferred to be grown thermally with a thickness of between about 3.6nm to 5nm, the silicon nitride layer deposited by CVD deposition is about 2 to 5 nm and the top oxide layer is deposited by CVD deposition and about 5 to 8 nm thick. The top oxide CVD layer is slightly thicker compared to the first and second process embodiments, for subsequent polysilicon and disposable sidewall spacer etch stop. Then the polysilicon 245 for gate material is deposited by CVD and followed by CVD silicon nitride 232 deposition thickness of between about 50 to 100 nm.

Then a photoresist layer is formed and a masking process with exposure and development to define memory gates 245 are performed. The polysilicon layer is now etched vertically by reactive ion etching (RIE), using the under layer top silicon oxide in the composite

layer 230 as a etch stop. Then extra boron 202 is ion implanted at low energy (less than 10 keV power and ion dosage of between about 5E12 to 2 E13 ions per cm^2, also shallow As is implanted at the same time at about 5E12 to 1.5E13 at the same KeV range as is the boron, as shown in Fig. 6A. Even though the channel threshold is very low due to As compensation, there is plenty of impurity to create a channel potential drop in the short channel region.

A thin silicon oxide layer 234 of about 5 nm is thermally grown on the side of polysilicon or CVD uniformly deposited. Then a disposable polysilicon layer typically having a thickness of between about 90 to 150nm is deposited. Then the vertical or anisotropic polysilicon etch is performed, which forms the disposable sidewall spacer 243 in Fig 6B. This is a thicker spacer than in the first and second embodiments. Then As ions are implanted at dosage of between 1E15 to 5E15 cm^2 and at the energy range of 20 to 50 KeV through the composite layer of oxide-nitride in order to form an N+ junction 204. By adjusting the lateral out diffusion with annealing temperature and time (between 850 to 900 ℃ and 5 to 20 min), the channel length defined from the edge of the word gate to the N+ junction edge is designed to be about 30 to 50 nm (3 to 4 times the electron mean free length) for ballistic high injection efficiency at low voltage.

Afterwards, the disposable side wall spacer 243 is gently removed by a dry chemical, isotropic etch. A typical etch ambient for this step is $HBr/CL_2/O_2$. The exposed silicon oxide over nitride is gently etched out by buffered hydrofluoric acid. A fresh silicon oxide 244 replacing the top oxide in the composite ONO 230, shown in Fig. 6C, of about 4 to 6 nm is deposited by chemical vapor deposition. Thermal oxidation is added after the top layer is deposited to improve the top oxide quality.

As an option, prior to removal of the disposable sidewall spacer 243, the exposed top two layers of oxide-nitride are etched by RIE. Then the fresh oxide of about 4 to 6 nm is deposited by chemical vapor deposition and followed by thermal oxidation for the top oxide improvement. During this oxidation process of about 850 to 900 °C and 10 min in wet O_2 atmosphere an extra oxide layer of about 20nm is formed on the nitride cut area over the n+ junction as shown by 244 in Fig. 6D. This thick oxide reduces the coupling capacitance between control gate 240 and bit diffusion 204.

A layer of polysilicon approximately 300 nm, which is slightly thicker than the summation of word polysilicon 245 and the top nitride 232 height, is deposited and CMP is performed using the nitride layer as the etch stop layer. Then the filled polysilicon layer 240 is recessed about 50nm by a vertical, anisotropic reactive ion etch. Then thin Ti or Co of about 10 nm is deposited and silicidation is performed. The silicide layer 241 is to reduce the control gate resistance. A CVD SiO_2 deposition and CMP is performed again, as illustrated by 236. The cross section of the device at this point is shown in Fig. 6C and in Fig. 6D.

Then the nitride layer 232 is selectively etched by H_3PO_4 or etched by a chemical dry etch. The polysilicon layer 248 having a thickness of between 150 and 200 nm is deposited by CVD. This polysilicon layer and underlying word gate polysilicon 245 are defined by normal photoresist and RIE processes. The structure at this point is as shown in Fig 6F.

The polysilicon layer 248 acts as a word line wire by connecting adjacent word line gates. The final memory cell is completed at this point. This word polysilicon layer can be silicided with Ti or Co to reduce the sheet resistance. A typical bird's-eye view of the memory cell is shown in Fig. 4G. The shallow trench isolation region is provided by the area 209. It is

understood that these critical dimensions will scale with the technology as the critical dim nsion is reduced.

In the embodiments described above, two approaches have been combined to improve memory density in this invention. In the first approach, density is more than doubled by sharing as many cell elements as possible. A single word select gate is shared between two nitride charge storage regions, and source lines/bit lines as well as control gate lines are shared between adjacent cells. In the second approach, multi-level thresholds are stored in the nitride regions under the control gates, and specific voltage and control conditions have been developed in order to make multi-level sensing and program possible for the high density array, with good margins between each of the threshold levels.

OPERATING METHOD FOR MULTILEVEL STORAGE

The procedures described below can be applied to multi-level storage of two bits or greater, as well as single-bit/two level storage applications in which Vt-hi and Vt-low are the highest and lowest threshold voltages, respectively, to be stored in the nitride region under the control gate. The dual bit nature of the memory cell comes from the association of two nitride regions paired to a single word gate and the interchangeability of source and drain regions between cells. This cell structure can be obtained by a side wall deposition process, and fabrication and operation concepts can be applied to both a step split ballistic transistor and/or a planar split gate ballistic transistor. The step split and the planar ballistic transistors have low programming voltages, fast program times, and thin oxides.

A cross-section of the array for a planar split gate ballistic transistor application is shown in Figure 7B. All word gates 340, 341, and 342 are formed in first level polysilicon and connected together to form a word line 350. ONO is formed underneath the

sidewalls that are deposited in pairs on either side of the word gates 340, 341, and 342. The nitride within the ONO layer which is under each sidewall is the actual region for electron memory storage. These nitride regions are 310, 311, 312, 313, 314,315 in Figs. 7B and 7C. In order to simplify peripheral decode circuitry, two side wall control gates sharing the same diffusion will be connected together to form a single control gate 330, 331, 332, 333, according to process embodiment 3 and embodiments 1 and 2 in which the gap-filling material 247 is a conductor. In the cases of process embodiments 1 and 2 in which two side wall gates sharing a diffusion are isolated from each other (where the gap-filling material is an insulator), it is feasible to electrically connect these two gates together with a wire outside of the memory array. Although it is also possible to operate the memory array with individual sidewall gates as control gates, peripheral logic will become more cumbersome, which does not meet the interests of high density memory.

Nitride regions 311 and 312 share control gate 331, and nitride regions 313 and 314 share control gate 332. A memory cell 301 can be described as having a source diffusion 321 and bit diffusion 322, with three gates in series between the source diffusion and the bit diffusion, a control gate 331 with underlying nitride region 312, a word gate 341, and another control gate 332 with underlying nitride region 313. The word gate 341 is a simple logical ON/OFF switch, and the control gates allow individual expression of a selected nitride region's voltage state during read. Two nitride charge regions which share the same word gate will be hereinafter referred to as a "nitride charge region pair". Within a single memory cell 301, one nitride charge region 313 is selected within a nitride charge region pair for read access or program operations. The "selected nitride charge region" 313 will refer to the selected nitride region of a selected nitride pair. The "unselected nitride charge region" 312 will refer to the unselected nitride charge region of a selected nitride charge region pair. "Near adjacent nitride charge regions" 311 and 314 will refer to the nitride charge regions of the nitride charge pairs in the adjacent unselected memory cells which are closest to the selected memory cell 301. "Far unselected adjacent nitride charge

regions" 310 and 315 will refer to the nitride charge regions opposite the near unselected adjacent nitride charge regions within the same unselected adjacent memory cell nitride charge region pairs. The "source" diffusion 321 of a selected memory cell will be the farther f th two memory cell diffusions from the selected nitride charge region and the junction closest to the selected nitride charge region will be referred to as the "bit" diffusion 322.

In this invention, control gate voltages are manipulated to isolate the behavior of an individual nitride charge region from a pair of nitride charge regions. There are three control gate voltage states: "over-ride", "express", and "suppress". A description of the control gate voltage states follows, in which the word line voltage is assumed to be 2.0V, the "bit" diffusion voltage is 0V, and the "source" diffusion voltage is assumed to be 1.2V. It should be understood that the voltages given are examples for only one of many possible applications, depending on the features of the process technology, and are not to be limiting in any way. In the over-ride state, the V(CG) is raised to a high voltage (~5V) forcing the channel under the control gate to conduct regardless of the charge stored in the nitride regions. In the express state, the control gate voltage is raised to about Vt-hi (2.0V), and the channel under the control gate will conduct, depending on the programmed state of the nitride regions. In suppress-mode, the control gate is set to 0V to suppress conduction of the underlying channels.

Table 1 gives the voltages during read of selected nitride region 313.

Voltages for Read of Selected FG=313

Vd0 320	Vcg 0 330	Vwl 340	Vd1 321	Vcg 1 331	Vwl 341	Vd2 322	Vcg 2 332	Vwl 342	Vd3 323	Vcg 3 333
0*	0	2.5	1.2	5	2.5	~0	2.5	2.5	0*	0

Table 1

*If threshold voltage is slightly negative, it is possible to suppress the nitride thresh ld region with a slightly negative control gate voltage (about –0.7V)

During read operation of nitride region 313, shown in Fig. 3C, the source line 321 can be set to some intermediate voltage (~1.2V) and the bit line 322 may be precharged to 0V. In addition, the following conditions must be met in order to read a selected nitride charge region: 1) the word select gate voltage must be raised from 0V to a voltage (2.5V) which is some delta greater than the sum of the threshold voltage of the word select gate (Vt-wl=0.5V) and the source voltage (1.2V), and 2) the voltage of the control gate above the selected nitride charge region must be near Vt-hi ("express"). The voltage of the control gate above the unselected nitride charge regions must be greater than the source voltage plus Vt-hi ("over-ride"). The control gates above the unselected adjacent nitride charge regions must be zero ("suppress"). The voltage of the bit diffusion 322 can be monitored by a sense amplifier and compared to a switch-able reference voltage, or several sense amplifiers each with a different reference voltage, to determine the binary value that corresponds to nitride charge region 313's threshold voltage, in a serial or parallel read manner, respectively. Thus, by over-riding the unselected nitride region within the selected memory cell, and then suppressing the adjacent cell unselected nitride regions, the threshold state of an individual selected nitride region can be determined.

For ballistic channel hot electron injection, electrons are energized by a high source-drain potential, to inject through the oxide and onto the nitride. The magnitude of the programmed threshold voltage can be controlled by the source-drain potential and the program duration. Table 2 describes the voltages to program multiple threshold voltages to a selected nitride region 313. These voltages are for example only, to facilitate description of the program method, and are not limiting in any way. In Table 2A, the control gates 331, 332 associated with

the s lected memory cell 301 are raised to a high voltage (5V) to over-ride th nitride charge regions 312 and 313.

Bit Diffusion Method Program of Selected Nitride Charge Region 313

Vt Data	Vd0 320	Vcg 0 330	Vwl 340	Vd1 321	Vcg 1 331	Vwl 341	Vd2 322	Vcg 2 332	Vwl 342	Vd3 323	Vcg 3 333
00	0	0	2.0	~0	5	2.0	5	5	2.0	0	0
01	0	0	2.0	~0	5	2.0	4.5	5	2.0	0	0
10	0	0	2.0	~0	5	2.0	4.0	5	2.0	0	0

Table 2A

Program of the desired threshold level is determined by the bit diffusion 322. The bit diffusion 322 is fixed to 5V, 4.5V, or 4.0V in order to program threshold voltages of 2.0V, 1.6V and 1.2V, respectively. When the word line 350 is raised above the word gate's 341 threshold, high energy electrons will be released into the channel, and injection begins. To inhibit program in the adjacent memory cells, the far adjacent control gates are set to 0V, so there will be no electrons in the channels of the adjacent memory cells. Thus, multi-level threshold program can be achieved by bit diffusion voltage control for this high density memory array. It is also possible to program multiple thresholds by varying the word line voltage, for example 4.5V, 5V and 5.5V, to program 1.2V, 1.6V and 2.0V, respectively.

Another possible method of program is to vary the control gate voltage in order to obtain different threshold levels. If multi-levels are to be obtained by control gate voltage, the unselected control gate 331 within the selected memory cell 301 will be set high to 5V in order to over-ride nitride region 312. The control gate 332 over the selected nitride region 313 will be varied to 4.5V, 5V and 5.5V, to obtain threshold voltages of 1.2V, 1.6V and 2.0V, respectively.

A fourth program method variation to the voltage conditi ns described for multi-level program is given in Table 2B, in which the selected control gate voltage matches the bit voltage for Vd=5V, 4.5V, and 4.0V and Vcg=5V, 4.5V, and 4.0V, respectively.

Control Gate-Bit Method Program of Selected Nitride Charge Region 313

Vt Data	Vd0 320	Vcg 0 330	Vwl 340	Vd1 321	Vcg 1 331	Vwl 341	Vd2 322	Vcg 2 332	Vwl 342	Vd3 323	Vcg 3 333
00	0	0	2.0	~0	5	2.0	5	5	2.0	0	0
01	0	0	2.0	~0	4.5	2.0	4.5	4.5	2.0	0	0
10	0	0	2.0	~0	4.0	2.0	4.0	4.0	2.0	0	0

Table 2B

Because the program current is low, and by programming schemes described above, it is possible to program several cells on the same word line in a parallel operation. Furthermore, depending on the peripheral decoding circuitry, multiple thresholds may also be programmed simultaneously, if the program methods of bit diffusion or control gate control are used. It should be noted however, that selected memory cells can have no fewer than two memory cells between each other, in order to obtain properly isolated behavior. Also, in order to obtain the tight Vt margins which are necessary for multi-level operation, the threshold voltage should be periodically checked during program, by a program verify cycle which is similar to a read operation. Program verify for the ballistic short channel sidewall MONOS in this invention is simpler than conventional floating gate and MONOS memories because program voltages are so low and very similar to read voltage conditions.

Removal of electrons from the nitride region during erase can be done by hot hole injection from the nitride region to the diffusion, or by F-N tunneling from the nitride region to the control gate. In the hot hole injection method, the substrate is grounded, diffusions are set to

5V and negative 5V is applied to the c ntrol gate. For F-N tunneling, a negativ 3.5V is applied to both th substrate and diffusions and positive 5V is applied to the control gates. A block of nitride regions must be erased at once. A single nitride region cannot be erased.

PREFERRED EMBODIMENT FOR READ

Read operation for a two bit multi-level storage in each of the nitride regions will be described, based on simulations for a 0.25u process. Figure 8A illustrates the memory cell and voltage conditions for a read of nitride charge region 313. The threshold voltages for the four levels of storage are 0.8V, 1.2V 1.6V and 2.0 for the "11", "10", and "01" and "00" states, respectively. This is shown in Figure 8B. The threshold voltage for the word select gate is 0.5V. During read, the source voltage is fixed to 1.2V. The control gate above the unselected nitride charge region is set to 5V, which overrides all possible threshold states, and the control gate above the selected nitride charge region is set to 2.0V, which is the highest threshold voltage of all the possible threshold states. All other control gates are set to zero, and the bit junction is precharged to zero. The word line is then raised from 0V to 1.0V, and the bit junction is monitored.

Sensing the bit junction yields the curves shown in Figure 8C. Bit line voltage sensing curves 71, 73, 75, and 77 during read of nitride charge region 313 are shown for different thresholds 0.8V, 1.2V, 1.6V, and 2.0V, respectively. It can be seen from the voltage curves, that the voltage difference between each of the states is approximately 300mV, which is well within sensing margins. Simulation has also confirmed that the sate of the unselected cell has very little impact on the bit junction voltage curve in Figure 8C.

The present invention provides a method for forming a double side wall control gate having an ONO nitride charge storage region underneath with an ultra short channel.

The enhancement mode channel is around 35nm, and is defined by the side wall spacer. The isolation between th word gates is formed by a self-aligned SiO_2 filling technique. The polysilicon control gate is formed by a self-aligned technique using chemical mechanical polishing. The process of the invention include two embodiments: a planar short channel structure with ballistic injection and a step split short channel structure with ballistic injection. A third embodiment provides isolation of adjacent word gates after control gate definition.

While the invention has been particularly shown and described with reference to the preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made without departing from the spirit and scope of the invention.

According to the present invention, a fast low voltage ballistic program, ultra-short channel, ultra-high density, dual-bit multi-level flash memory is achieved with a two or three polysilicon split gate side wall process.

4. Brief Explanation of the Drawings

Figure 1 is a device structure of prior art SONOS (Silicon Oxide Nitride Oxide Silicon).

Figure 2A graphically represents empirical results for a split gate floating gate transistor, demonstrating that for a channel length of 100nm, source side injection requires high voltage operation.

Figure 2B graphically represents empirical results for a split gate floating gate transistor showing that for a channel length of 40nm, ballistic injection operates at much lower voltages and/or much faster program speed.

Figure 3A is an array schematic of the prior art double side wall dual-bit split floating gate cell with ultra short ballistic channel.

Figure 3B is a layout cross-section of the prior art double side wall dual-bit split floating gate cell with ultra short ballistic channel.

Figs. 4A through 4F are cross sectional representations of a first preferred embodiment of the process of the present invention.

Fig. 4G is a bird eye's view of the completed memory cell of the present invention.

Figs. 5B, 5C, and 5F are cross sectional representations of a second preferred embodiment of the process of the present invention.

Figs. 6A through 6F are cross sectional representations of a third preferred embodiment of the process of the present invention.

Figure 7A is an array schematic of the present invention.

Figure 7B is a cross-sectional representation of the present invention.

Figure 7C gives the required voltage conditions during read for the present invention.

Figures 8A, 8B, and 8C are graphical representations of voltage sensing curves for the present invention during read.

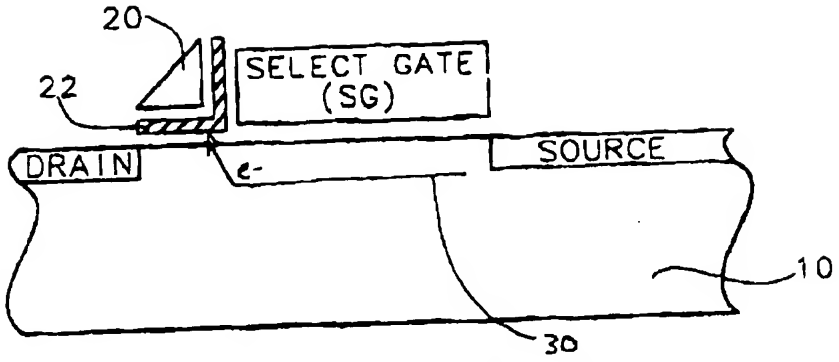


FIG. 1 Prior Art

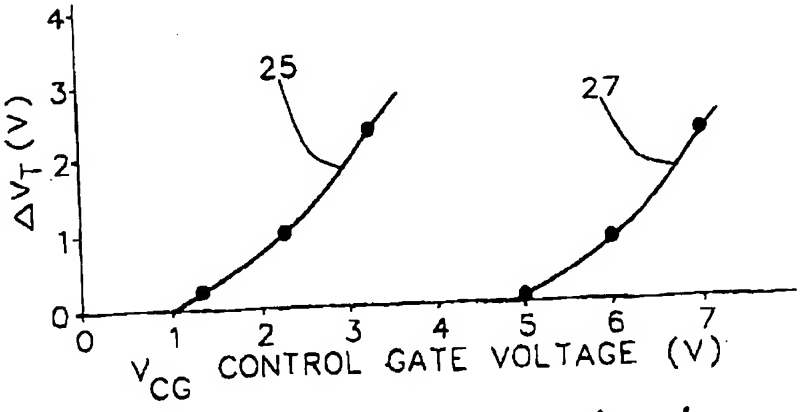


FIG. 2A Prior Art

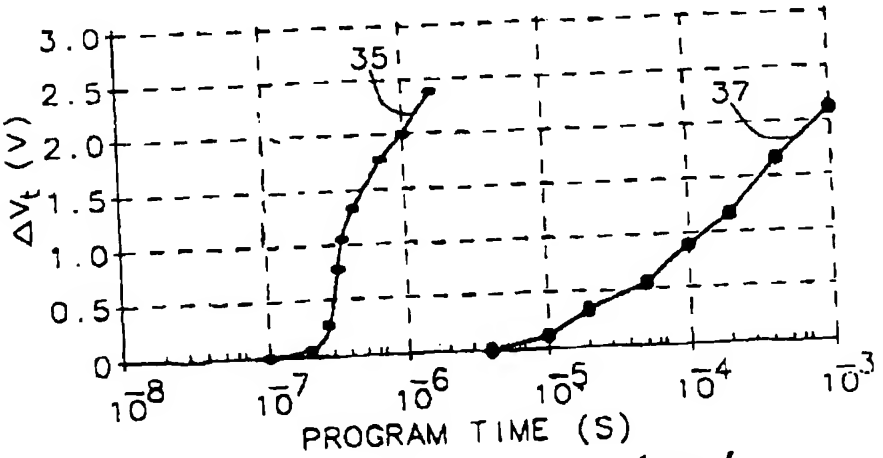


FIG. 2B Prior Art

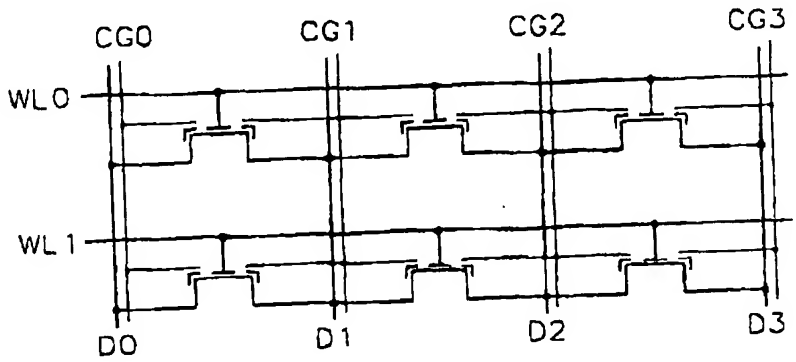


FIG. 3A Prior Art

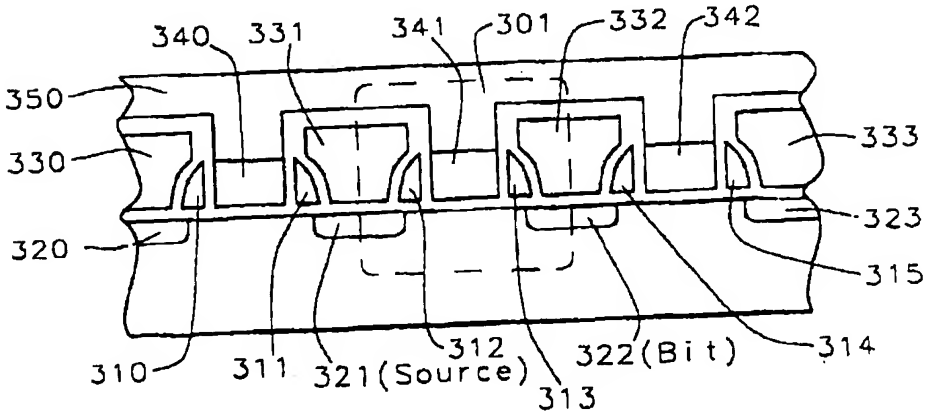


FIG. 3B Prior Art

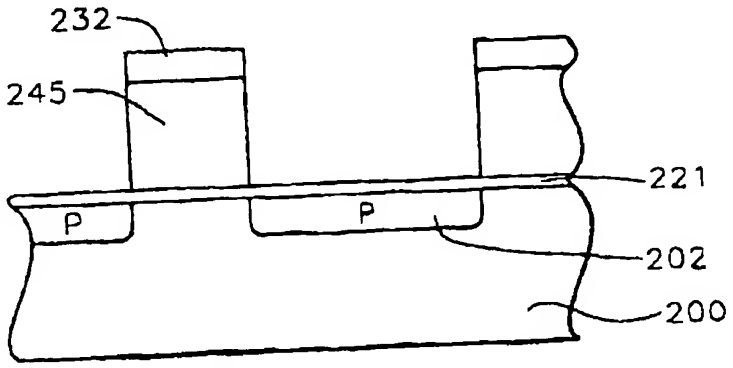


FIG. 4A

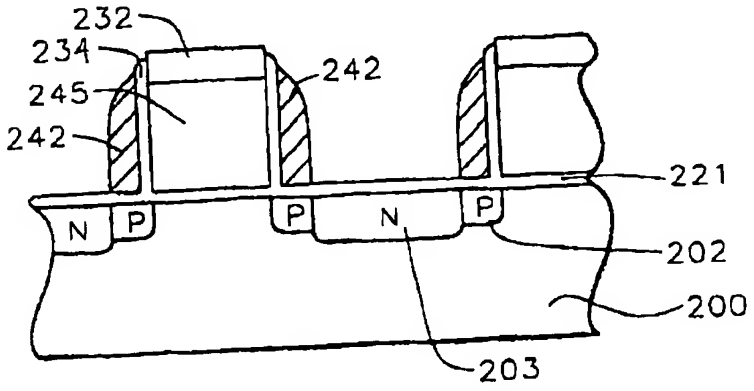


FIG. 4B

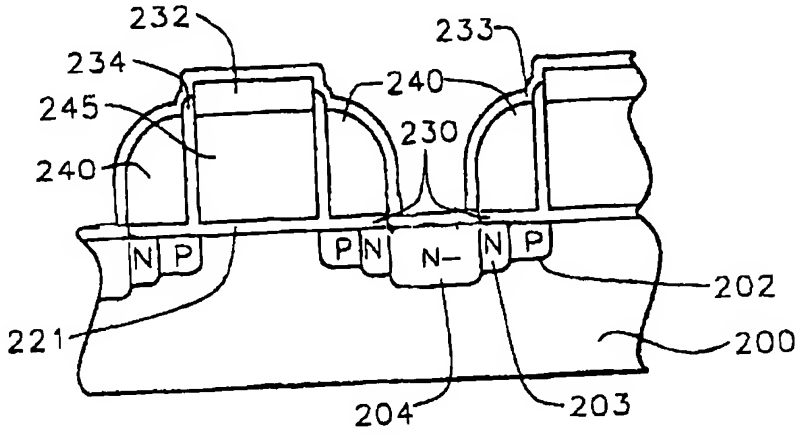

232
233
234
240
245
240
230
N
P
P
N
N-
N
P
202
221
200
204
203

FIG. 4C

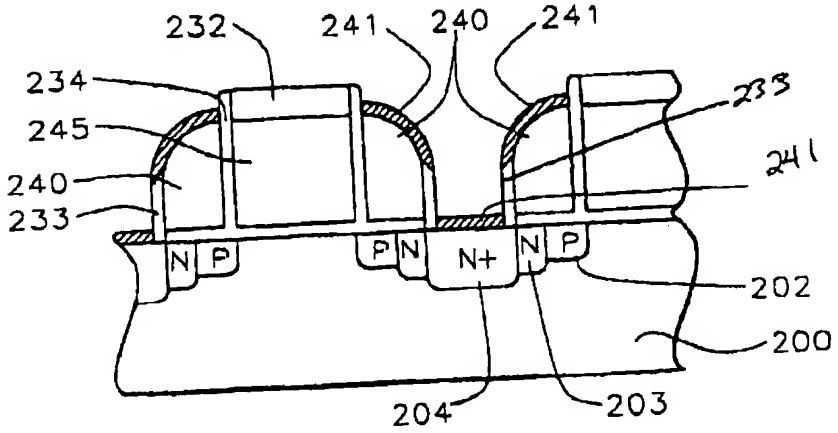

232
241
240
241
234
233
245
241
240
233
N
P
P
N
N+
N
P
202
200
204
203

FIG. 4D

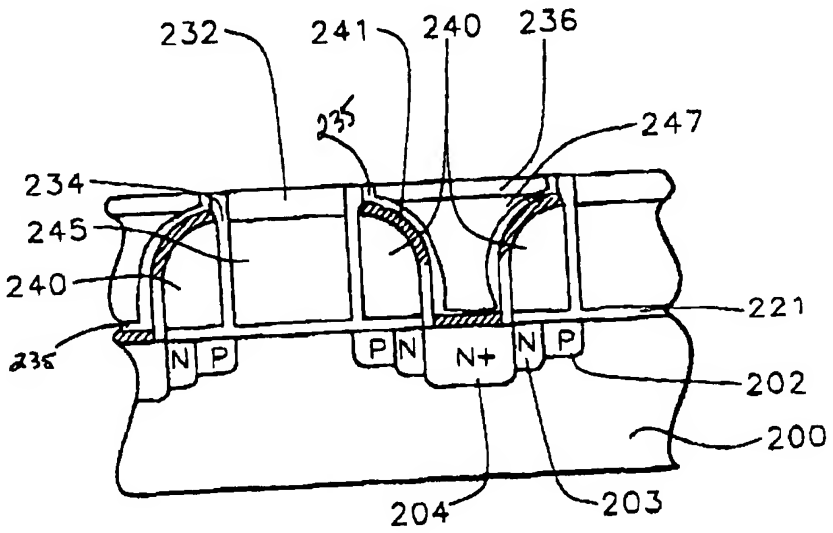

232
241
240
236
235
247
234
245
240
221
235
N
P
P
N
N+
N
P
202
200
204
203

FIG. 4E

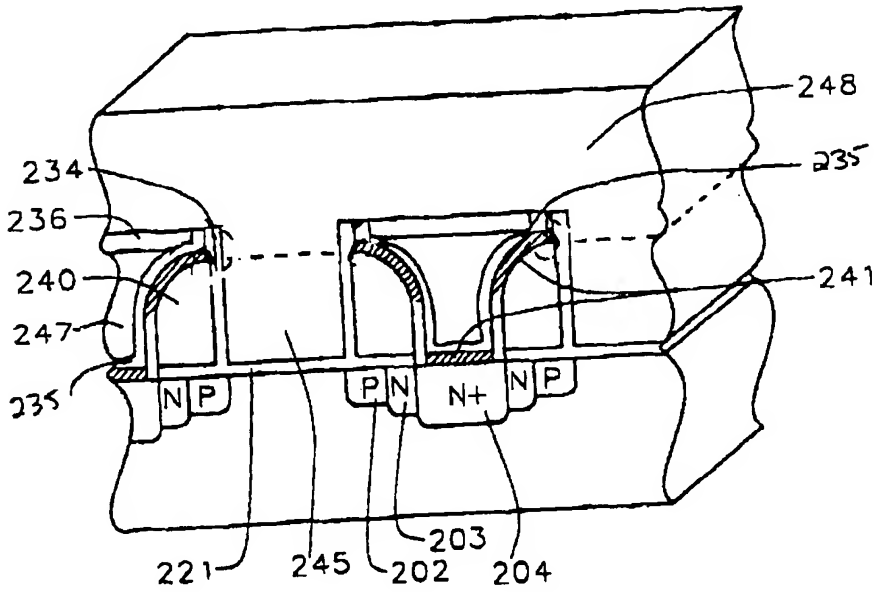

248
234
235
236
241
240
247
235
N
P
P
N
N+
N
P
221
245
202
203
204

FIG. 4F

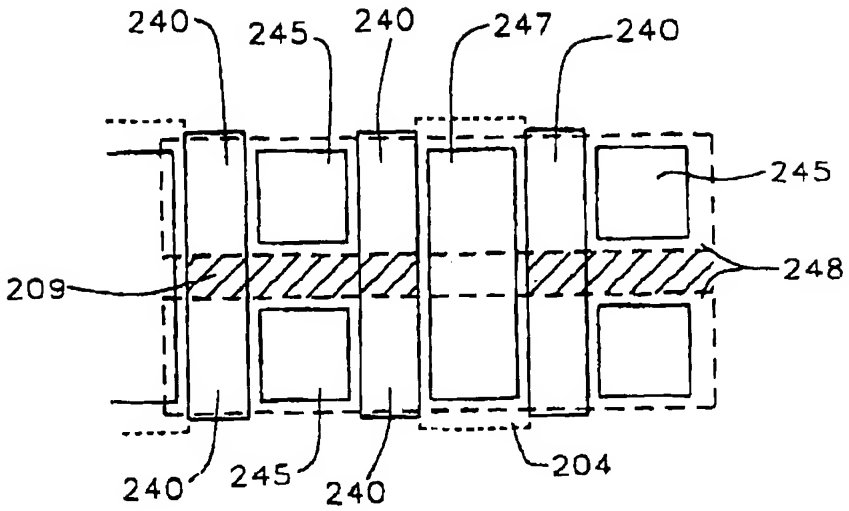

240
245
240
247
240
245
248
209
240
245
240
204

FIG. 4G

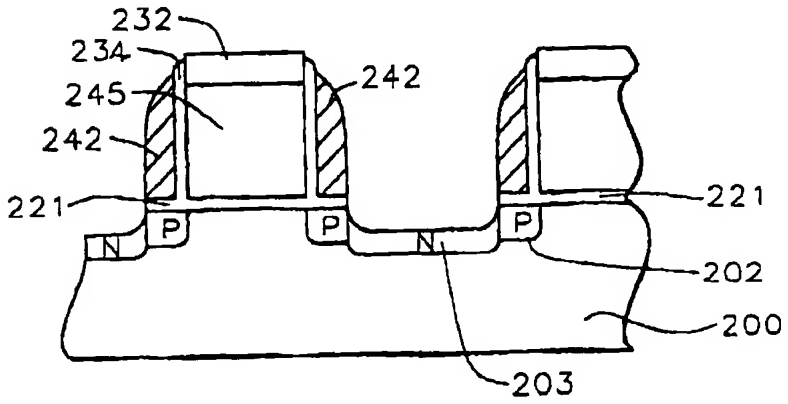

232
234
245
242
242
221
221
P
P
P
N
N
202
200
203

FIG. 5B

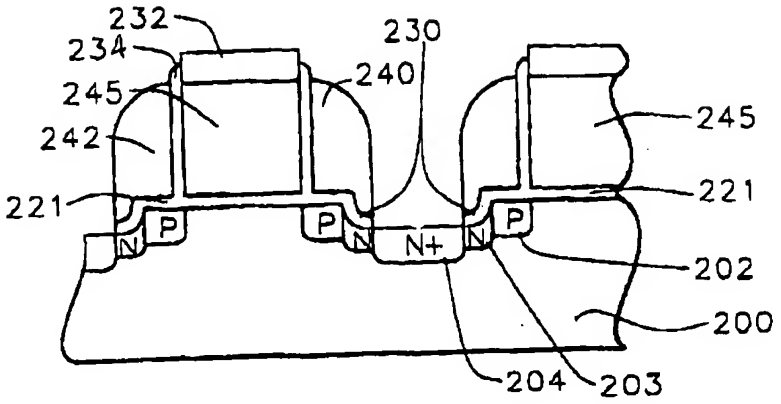

232
234
245
242
221
230
240
245
221
202
200
204
203
P
N
N+

FIG. 5C

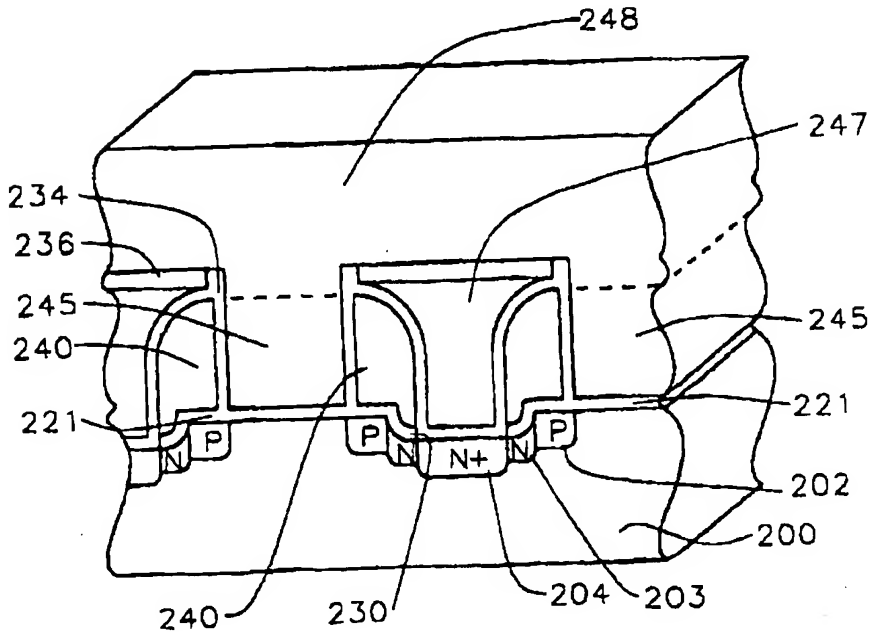

248
247
234
236
245
240
221
245
221
202
200
240
230
204
203
P
N
N+

FIG. 5F

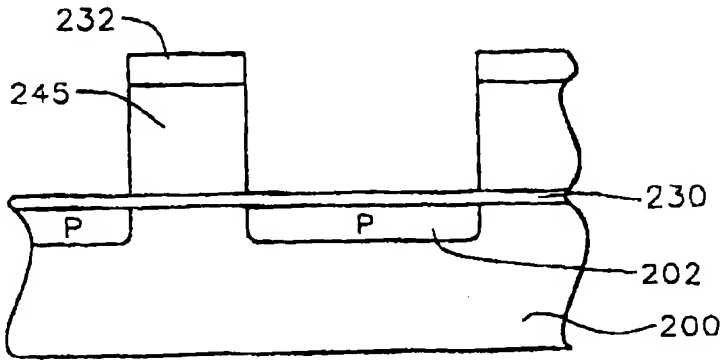


FIG. 6A

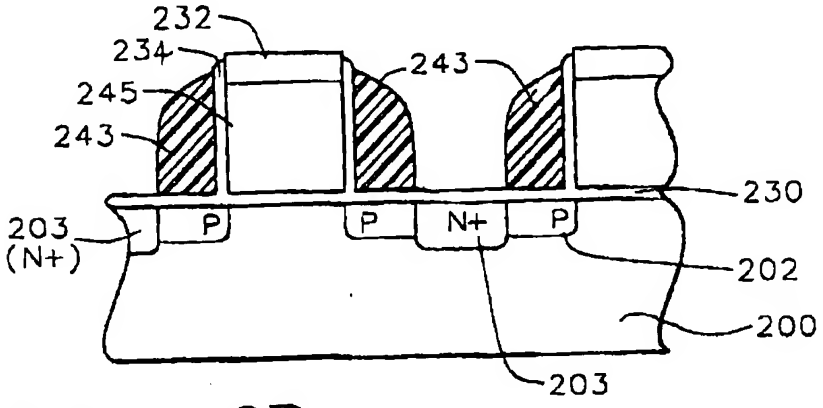


FIG. 6B

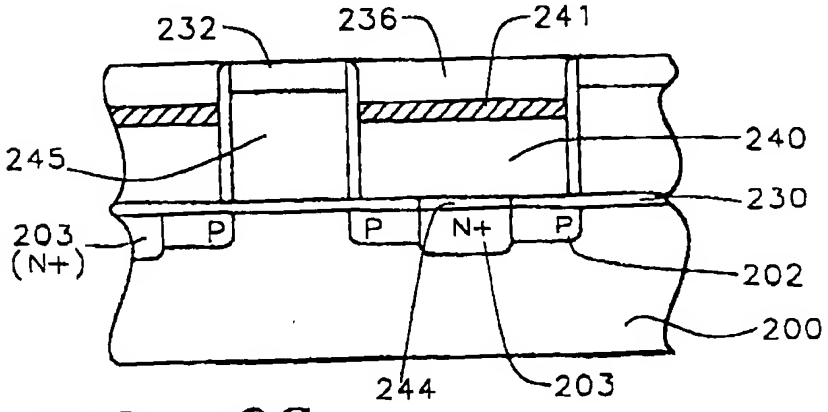


FIG. 6C

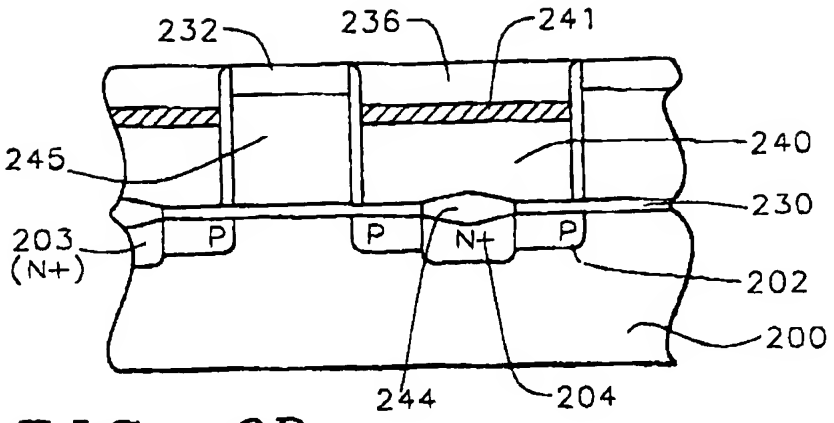

232
236
241
245
240
230
203
(N+)
P
P
N+
P
202
200
244
204

FIG. 6D

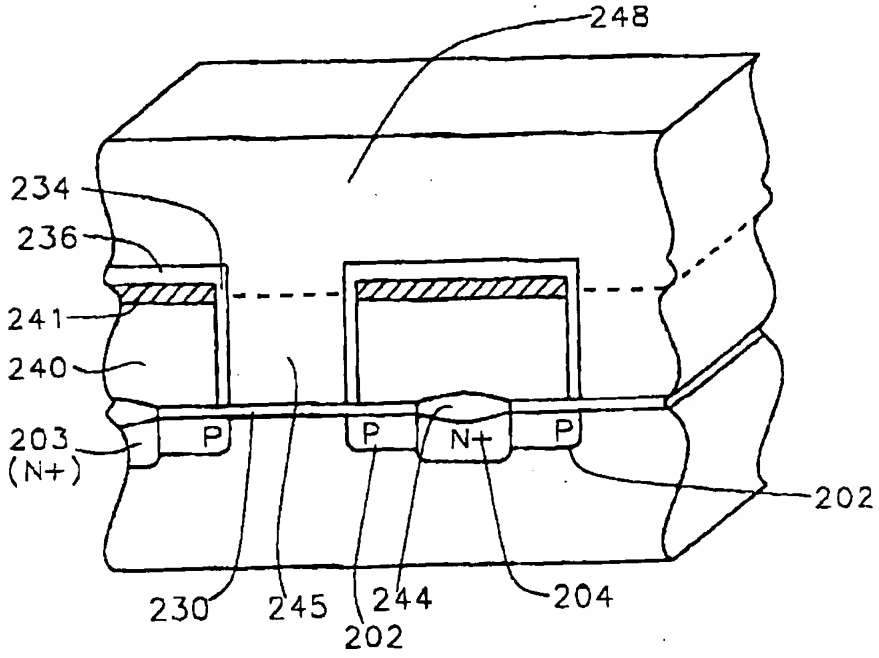

248
234
236
241
240
203
(N+)
P
P
N+
P
202
230
245
244
202
204

FIG. 6F

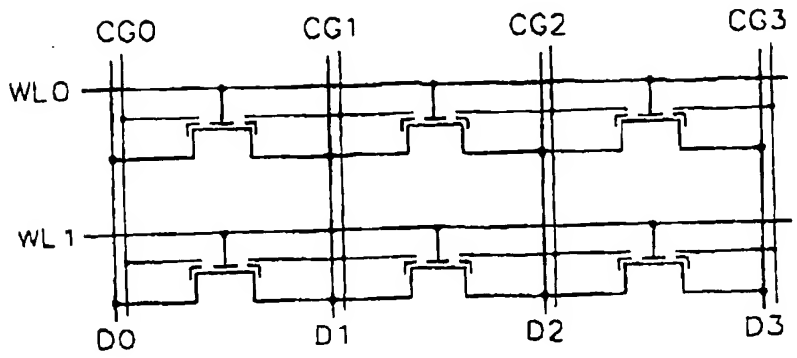


FIG. 7A

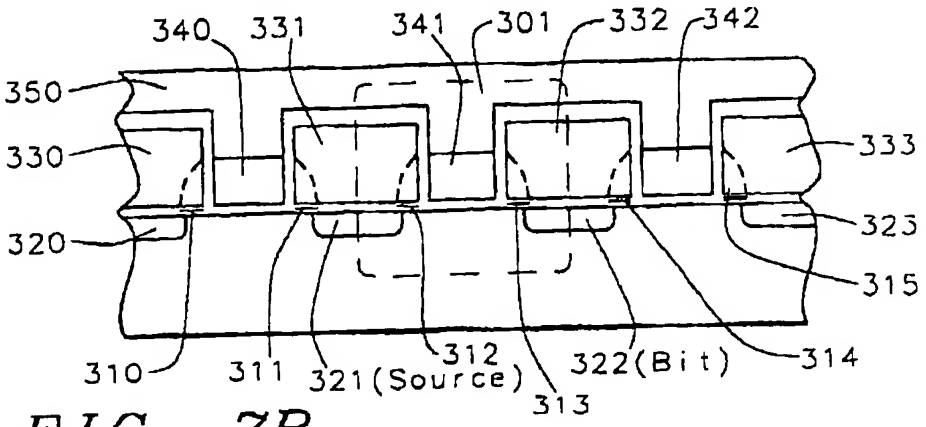


FIG. 7B

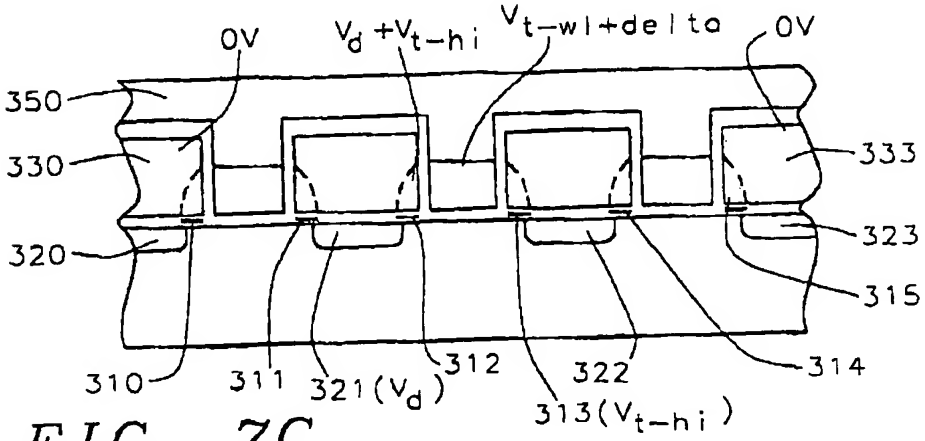


FIG. 7C

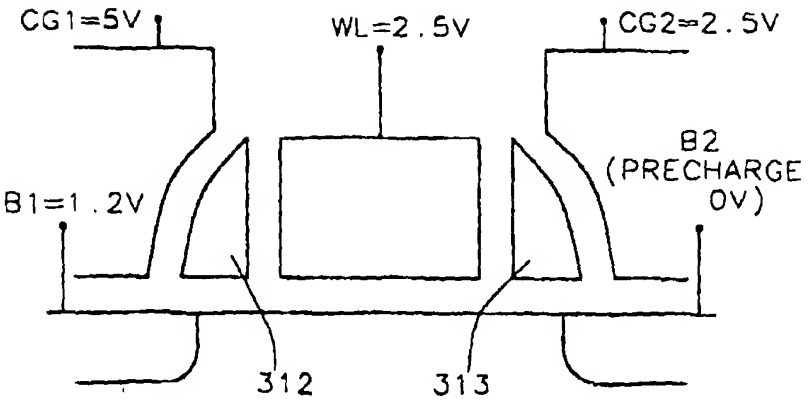


FIG. 8A

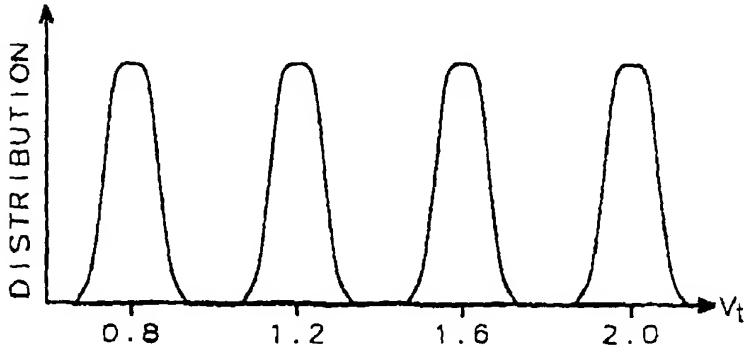


FIG. 8B

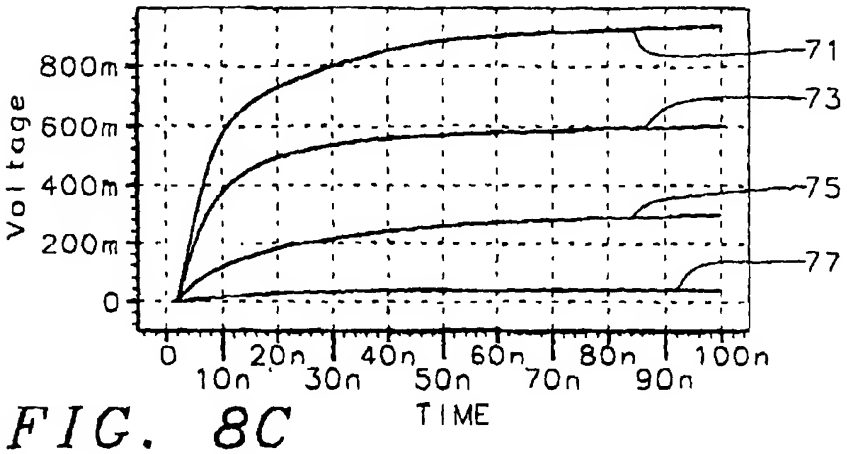


FIG. 8C

1. Abstract

In this invention, a fast low voltage ballistic program, ultra-short channel, ultra-high density, dual-bit multi-level flash memory is disclosed with a two or three polysilicon split gate side wall process and it operation. The structure and operation of this invention is enabled by a twin MONOS cell structure having an ultra-short control gate channel. The cell structure is realized by (i) placing side wall control gates over a composite of Oxide-Nitride-Oxide (ONO) on both sides of the word gate, and (ii) forming the control gates and bit impurity layer by self-alignment and sharing the control gates and bit impurity layers between neighboring memory cells for high density. Key elements used in this process are: 1) Disposable side wall process to fabricate the ultra short channel and the side wall control gate with or without a step structure, and 2) Self-aligned definition of the control gate over the storage nitride and the impurity layer.

2. Representative Drawing

Fig.4F